THE EFFECTS OF FAILURE BY THREE EVALUATION METHODS ON THE ATTITUDES OF STUDENTS

By

BRUCE DALE COOK

Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
1973

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1976

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF EDUCATION
July, 1979

THE EFFECTS OF FAILURE BY THREE EVALUATION METHODS ON THE ATTITUDES OF STUDENTS

Thesis Approved:

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGMENTS

The author wishes to express his appreciation to his committee chairman and major adviser, Dr. John Hampton, for his guidance and expertise in critical thinking. Appreciation is also extended to committee members Dr. Paul Warden, Dr. David Perrin, Dr. Russell Dobson, and Dr. Darrell Ray for their time and assistance given to this research, and to the author's personal, academic, and professional growth.

A note of special thanks is extended to Julie Cook and Kathy Paige for their assistance in typing the earlier drafts and final copy of the manuscript. Sincere appreciation is also expressed to Paul and Wanda Cook, and to George and Mildred Rizek for their continual encouragement. Finally, special gratitude is expressed to my wife, Julie, for her understanding, encouragement, and many personal sacrifices over the past two years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

THE RESEARCH PROBLEM

Nature of the Problem

Included in the educational system are concerns about evaluating the extent to which a student has learned the material taught, that is the manifest curriculum (Bloom, 1976). Although there is much criticism of grades as a means of evaluation, they are widespread in their use as measures of the quality of learning and quantity learned.

The assignment of grades has traditionally been based on the use of relative evaluations which consists of assessing a student's performance in terms of his relative standing to a group of peers (norm-references). This evaluation method has typically made use of percentages based upon the normal curve or some variation of that theme (Terwilliger, 1973). The typical student experiences years of being judged in relation to others by the comparisons made between classmates or larger reference groups (e.g. state or national norms) and one's own performance. This comparison yields some measure of relative standing; that is, if achievement is viewed as a continuum from most to least, then "Where on that continuum does a particular student fall?" Grades have often been assigned on the basis of the answer to that question. The problem of this method in evaluating students' performance is that the students become "locked-in" to a relative standing.

From one learning task to a related learning task the individual student's performance generally increases, but so do the performances of the other students (the norm). So, while a particular student has performed better than last time, that student's relative standing remains somewhat the same since most other students also performed better. Since relative standing is somewhat resistant to change, a student comes to know his place in a group regardless of his actual performance level, that is the latent curriculum (Bloom, 1976).

It is the student's perception of his performance that is important in evaluation. With norm-referenced evaluation the student's perception is based on the relative judgment of others. This type of evaluation has met with criticism in terms of its arbitrariness in deciding grades, not encouraging individualized learning, and being poor in predicting success in addition to producing excessive competition, negativeness, and dishonesty (Astin et al., 1967). Others are now criticizing the norm-referenced approach in terms of the affective outcomes of relative evaluations.

Absolute measures (criterion-referenced, competency testing, mastery learning) are being suggested to correct many of the defects of relative measures (Block, 1971; Block, 1974; Bloom, 1971; Carroll, 1963; Popham, 1969; Proger & Mann, 1973). With this type of evaluation, student performance is judged on some criterion of adequacy which is, in the student's perception, independent of his relative standing in a group of peers. In other words, there is a specific level that has been established to represent acceptable performance on a learning task. Unlike norm-referenced methods which require students to be

spread out on a continuum, the criterion-referenced methods allow for the possibility that all students may achieve the same level of performance, and therefore the same grade.

Two general types of criterion-referenced evaluation methods can be postulated. The typical and most often used type is that of a regulated criterion level. With this type, there is an established and fixed minimal level of competence which has been arrived at by some authority figure (e.g. teacher). The decision to use that criterion level of acceptable performance is based on the opinion, intuitively or empirically formed, of the authority figure as to what constitutes mastery of that particular learning task. The second type of criterion-referenced evaluation can best be viewed as a negotiated form in which the acceptable level of competence is mutually established and agreed upon by both the authority figure and the student being evaluated. This requires that the student become an active member of his own evaluation by assisting in the setting of a specific competence level for a particular learning task.

The cognitive outcomes of criterion-referenced evaluation have met with considerable research, and the attitudinal outcomes, while not being ignored, have received less emphasis. Khan (1969) states

> . . . that the average relationship between aptitude variables . . . and achievement criteria ranges between .50 - .75 . . . one-half to three-quarters of the variability in academic achievement remains unexplained. Therefore, research on academic prediction has shifted toward the measurement of non-intellective factors in academic performance (p. 216).

As part of this shift toward researching the affective/attitudinal influence on performance (and vice versa), Bloom (1976) has suggested that success or failure in an absolute sense does not have the

pronounced effect on a student's affect as it would with relative evaluation. Bloom's belief remains to be validated, but it again brings to the forefront a continually asked question of educators:

> How can evaluation be incorporated into education so that learning is enhanced and positive student attitudes toward the learning process are developed? (Peckham & Roe, 1977, p. 41).

Since failure to perform apparently has a strong negative effect on the individual's attitudes and, logically, repeated failure should discourage learning, it becomes important to use an evaluation method that produces the least severe effects after failure. The question arises: "What effect does failure as evaluated by a norm-referenced, a regulated criterion-referenced, and a negotiated criterion-referenced measure have on the attitude toward self and the learning task?"

Purpose of the Study

Over the years different types and various forms of evaluation methods have been suggested for specific purposes or for general use. Most attention given to evaluation has concerned itself with the reliability and validity of a specified method in assessing some ability. In the past, if the effects of an evaluation method on the person being evaluated was of importance, it generally was so in terms of "Does it lead to improved cognitive or psychomotor performance?" Questions such as "How did the student like what was learned?" or "How did the student feel about the process of learning?" have received less emphasis. If Bloom (1976) is correct in his research, then approximately 20 to 25 percent of the variation in achievement can be accounted for in the student's affect which is one component of attitude, and that is too large

of a percentage to ignore.

The effects of achievement on attitudes and self-concept have received attention, but little attention has been given to the effects of different evaluation methods on attitudinal variables. The effects of the evaluation methods on attitudes in the proposed study could help to promote understanding of how evaluation effects students' interest and self-concept in relation to performance situations. Thus, this proposed study is seen as continuing the research investigating the complex relationship between learner, task, performance, and attitude.

Definition of Terms

Criterion-referenced evaluation: an individual's performance is compared to an established level of proficiency. Such tests are constructed "to yield measurements that are directly interpretable in terms of specified performance standards," (Glaser & Nitko, 1971, p. 653).

a) Regulated criterion-referenced: this evaluation method involves a predetermined level of proficiency which is established by someone other than the individual being evaluated.

b) Negotiated criterion-referenced: this evaluation method establishes a level of proficiency which is based on the negotiations between the individual evaluated and the individual conducting the evaluation. In other words, the level of proficiency is arrived at by mutual agreement.

Failure: a judgmental term indicating that an individual has not achieved a specific goal. In criterion-referenced evaluation acceptable performance is specified in performance terms. In norm-referenced evaluation failure is defined as to how one compares to others, and in

the present study this was specified in percentiles.

Level of aspiration: an individual's immediately desired goal (usually conceived as a step or series toward an ultimate goal), and "not his prediction of the actual outcome . . . nor his best imaginable performance," (Diggory, 1966, p. 139).

Norm-referenced evaluation: evaluation designed to measure a person in relation to a normative group (see Popham, 1971). The frequencies of scores of an identified group are used as the comparative standard so that an individual's performance can be interpreted in terms of where that individual falls within the group (percentile rank).

Probability of success: an individual's estimate of his ability or "power" to achieve a goal during one of the remaining trials. "It does not refer to the likelihood that he will make it on any particular trial, nor to the likelihood that he will achieve some private LA (goal) on the next trial" (Diggory, 1966, p. 137).

Self-evaluation: an individual's estimate of how good he thinks a specific ability of his is. More precisely, it is the individual's estimate of his own value or "goodness" as an individual who needs a specific ability. This definition excludes "global self-evaluation" which, it is assumed, requires extensive general feedback about numerous abilities over long periods of time to effect it.

Task-related affect: conceptually defined as an individual's liking and desire about and for a specific task; operationally defined as the extent that an individual wishes to voluntarily engage in additional tasks of the same type (see Bloom, 1976).

Assumptions and Limitations

It is assumed that the four dependent variables (level of aspiration, probability of success, self-evaluation, task-related affect) are continuous in nature and that they may be measured linearly in varying intensity of degree. Spearman's rho was computed on pre- and post- measures of self-evaluation and task-related affect, and on select pairs of trials in each evaluation group for probability of success and level of aspiration in an attempt to estimate reliability of the one item scales. Most of the achieved correlations were low suggesting a limitation to interpretation of the results. Further discussion can be found in the Instruments section of Chapter III.

Although attitude may be effected by various factors, it is assumed that the major factor influencing the learner's attitude is his perception of his competence with a task. His perception may be realistically accurate or biased.

Due to the nature of the task which the subjects performed and its limited general exposure to students, it is assumed that all subjects will approach the task with neutral affect, i.e. neither positive nor negative feelings in relation to it.

CHAPTER II

REVIEW OF LITERATURE

Introduction

Very little empirical research has been conducted on the attitudinal outcomes of various evaluation methods. Most of the research has been more general in nature, concerning itself mostly with attitudes after success or failure without concern toward how different definitions of success or failure themselves affected those attitudes. The theoretical and empirical information on the subject is presented in this chapter as a foundation for the present study. In Chapter I, three evaluation methods were presented (norm-referenced, regulated criterion-referenced, negotiated criterion-referenced), and while they are not the only possible methods available, they are used more frequently in schools than others. These three methods will be elaborated on in this chapter and discussed in terms of the four dependent variables (task-related affect, level of aspiration, probability of success, self-evaluation). The literature will be reviewed in sections for each of the four variables in which relevant material will form the foundation for each hypothesis. All eight hypotheses will be stated in the null. However, before proposing hypotheses on the evaluation methods and dependent variables, a brief discussion of attitudes is presented to establish a common frame of reference.

Attitudes

Attitudes are generally regarded as likes and dislikes; however, the likes and dislikes are not really attitudes, but are the evaluative responses resulting from attitudes. The attitudes themselves are unobservable, but have been quantitatively inferred in many ways including scale ratings, opinion statements, and behavioral changes. These types of measurements are used to assess the three broad areas of attitudes (Bem, 1970; Zimbardo & Ebbesen, 1970). The first area is the affective component which consists of the individual's liking of or emotional response to something. The task-related affect variable and self-evaluation variable in the current study are comprised mostly of this affective component of an attitude.

The second area of an attitude is the cognitive component which involves the individual's beliefs about or factual knowledge of something. In the present study, the level of aspiration variable and the probability of success variable are both primarily cognitive components of an attitude.

The third component of an attitude is behavioral and it includes the individual's overt behavior directed toward something. This component is not included in the present study.

One of the current theories on attitude aquisition is the self-perception theory of attitude follows behavior (Bem, 1967; Bem, 1968; Bem, 1970). This theory proposes that an individual partially relies on the same external cues (behavior) in identifying ones own internal state as others use to infer his internal state. In other words, an individual might use his school performance as a means to formulate his

attitude towards school and himself. Therefore, attitude toward something should be affected if the behavior is changed.

Task-Related Affect

There have been a number of studies exploring the relation between achievement and school-related affect (Khan, 1969; Khan & Weiss, 1973; Kurtz & Swenson, 1951; Malpass, 1953; Michael et al., 1964; Russell, 1969). The studies indicate that there is a relation between school-related affect and achievement. Much less research has been conducted on specific learning tasks and achievement in an attitude framework.

Bloom (1976) provides a summary of research on the relationships between achievement and subject-related affect (the student's interest in the subject or desire to participate in additional tasks of the same type). He reports correlations generally between .20 and .40 which are most clear for the extreme levels of achievement. Bloom believes these relationships are causal and influenced by the students' perceptions of adequacy or inadequacy on specific tasks. The more adequate they perceive themselves to be on a task, the more the task will be desired. The reverse outcome is postulated for perceptions of inadequacy; if the student perceives himself as doing less than adequate performance, he will increasingly dislike the task and desire to avoid it in the future if given a choice.

Bloom (1976) is speaking strictly about the students' perceptions of adequacy as based on the judgment of their performance by teachers and peers in relation to the performance of other students (norm-referenced evaluation). He implies this is not the case with criterion-

referenced evaluation when he speaks of norm-referenced evaluation as a system ". . . independent of success or failure in any absolute sense. It is dependent on local definitions of success or failure relative to other students in the class or school" (p. 149).

This division between the possible effects of norm-referenced and criterion-referenced evaluation appears plausible when one realizes that norm-references generally separate students on an ability while criterion-references focus on minimal competencies of an ability. According to Simon (1976, p. 74), norm-referenced evaluation " . . . asks the student to place himself or herself in the hands of the teacher for rewards based not on what he or she learned, but on whether others learned more."

Cartwright's (1942) induced failure on activities was based on a norm-referenced evaluation. After the subject completed the task, he was informed of his elapsed time and told that he had taken longer to complete the task than any other subject. When rating the activities for attractiveness, the subjects who failed generally reduced their rating of attractivness for the task failed and for tasks similar to it.

The inherent difficulties of "bettering" one's relative standing in a group should further complicate the negative effects of arbitrariness in defining adequacy. Simply performing better on a task than last time is not enough. The student must improve his performance, but he must do it at a higher rate than those in the norm-group if he is to improve his relative standing. If the rate of improvement of the norm group members is similar, then it would appear to the individuals that

their performance is the same as past performances although their actual level of competencies increased. For those students at the lower end of the norm group, it should be rather frustrating since they are continually seen as inadequate (as judged relative to others) eventhough they are improving. The perception of objective improvement would appear to be clouded by the perception of adequacy based on relative judgment.

On the other hand, Gerwirtz (1959) also found failure to affect the choice of subsequent tasks, but the evaluation of performance was based on a criterion-reference measure, that is, failure was defined as the inability to complete a puzzle. Some subjects gave up spontaneously, while others, after fumbling for a long time, were told to start over. The results indicated that failure lead to a reduction in their preference to play with puzzles similar to the one failed.

The criterion-referenced evaluation method should give more hope and encouragement to students since they are judged in terms of whether they reach a minimum level of adequacy. The student who fails to reach the criterion should be able to see objective improvement toward the criterion not dependent on the performance of others. While failure on a series of tasks under either norm- or criterion-referenced evaluation should produce a lower task-related affect than task-success would produce, the failure on norm-referenced evaluation should produce lower task-related affect than failure on criterion-referenced evaluation.

Both of the above evaluation methods can be viewed as involving external evaluation; that is, the student who is being evaluated has no input into the standards of evaluation. As long as the student's adequacy is being judged by external sources such as parents, teachers, and

peers, then the student feels threatened and belittled since the external evaluator has access to only a small portion of the relevant information (Wilhelms, 1967). The student begins to see himself at the mercy of others -- someone to be manipulated (cf. DeCharms, 1971).

It is possible to imagine an evaluation system which would actively engage the evaluated student in the establishment of the evaluation standards. Such a system might be conceived as an interactive or negotiated evaluation (Combs, 1976; Combs, 1963; Wilhelms, 1967). Both the student and teacher would establish the standards of acceptable performance. Since the student would perceive himself in more control of his evaluation, he should feel better about performing toward those standards. The student, as part of the evaluation process, would be more inclined to use not only the objective information available to an external evaluator, but also he should feel that the subjective information played a larger part in the judgment of adequacy or inadequacy. Therefore, the student could interpret "failure" in light of this information as a tolerable inability to achieve a goal rather than as being an inadequate person. Because the individual is part evaluator, he could view his "failure" as a temporary set-back to a goal rather than as being powerless to achieve the goal. It would appear that the failing individual would be more inclined to desire another chance to attempt success with the task in a negotiated situation.

The following null hypothesis is based on the above information:

1) There are no differences between the three evaluation methods on level of task-related affect.

Level of Aspiration

Research involving level of aspiration conceived this concept as a means to study a subject's expectation or goal for some future attempt on a task (see Diggory, 1966, p. 115-128, for a historical review). Its use on tasks of criterion-referenced evaluation has generally indicated that it rises as long as the individual's performance rises and lowers as performance lowers (Diggory & Morlock, 1964). The overall trend is for the level of aspiration to reflect the shape of the performance curve. Typically, the subject will set a fairly high initial level of aspiration, but will drastically reduce it after the first trial due to the feedback of falling quite short of their expectations. The level of aspiration then will continue to rise on each trial as long as performance increases. There should be significant differences between early adjacent trials as opposed to later trials with a negatively accelerated performance curve.

The following null hypothesis is formulated on the preceding information:

2) There are no differences between the ten trials on level of aspiration.

The effects of norm-referenced evaluation on level of aspiration has apparently not been studied. It can be reasoned, however, that level of aspiration would differ under norm-referenced evaluation by typically maintaining a higher level than with criterion-referenced evaluation. This should occur because the individual under norm-references will soon realize that in order to progress in relative standing, he must perform at a higher rate than the norm, and therefore,

he should set higher level of aspirations on the average than the criterion-referenced individuals. No appreciable difference in level of aspirations are expected between the regulated and negotiated criterion-referenced groups.

The following null hypothesis is based on the above material:

3) There are no differences between the three evaluation methods on level of aspiration.

The literature indicates that level of aspiration basically follows the shape of the performance curve; and therefore, no interaction effect between evaluation method and trials is hypothesized.

The above discussion leads to the following null hypothesis:

4) There are no interactions between evaluation methods and trials on level of aspiration.

Probability of Success

Researchers as summarized by Diggory (1966) have found that successful and unsuccessful performances on a task as evaluated with criterion-references are highly related to the individual's estimation of his probability of succeeding on the task. As long as the rate of improvement is in the direction of eventually meeting the criterion, then the individual's probability of success will remain high. On the other hand, when it appears obvious to the individual that he will not achieve the criterion, then probability of success drops (Feather, 1965). It is possible for level of aspiration to rise and probability of success to fall at the same time. This is characterized by the individual who continues to improve and sets ever increasing levels of

aspiration, but at the same time becomes increasingly discouraged about achieving the fixed criterion. In this case the performance curve begins to level off (negatively accelerating) so that there is less chance of reaching the criterion before the deadline. Therefore, it is expected that significant differences between trials should occur in probability of success estimates.

The following null hypothesis is presented based on the above discussion:

5) There are no differences between the ten trials on probability of success.

Since the negotiated criterion-referenced group is more involved in the establishment of the criterion goal, it is expected that its probability of success would be higher than the probability of success of the regulated group.

Again, as with level of aspiration, probability of success has not met with research from a norm-referenced stance. Since an individual evaluated by the norm must do more than simply increase performance, that is, he must increase at a higher rate than the norm, it should be more frustrating than with a criterion-referenced system. This frustration is expected to manifest itself in lower probability of success for the norm group than either criterion group.

This review forms the basis for the following null hypothesis:

6) There are no differences between the three evaluation methods on estimates of the probability of success.

The trend for probability of success in criterion-referenced evaluations with a negatively accelerated performance curve should drop

after the first trial but then increasingly rise until the last trial where it is expected to drop. As for the norm-referenced method, probability of success is also expected to drop after the first trial and to slowly increase thereafter until the midpoint of trials where it should begin to drop. This should occur because subjects in the norm-referenced group will become discouraged earlier than the criterion-reference subjects because of a lack of progress in percentile. There is also expected to be a drastic drop before the last trial in the norm-referenced method.

The following null hypothesis is based on the above information:

7) There are no interactions between evaluation methods and trials on probability of success.

Self-Evaluation

A number of studies indicate that self-concept and achievement in school is related (Alvord et al., 1967; Brookover et al., 1964; Centi, 1965; Diller, 1954; Gibby & Gibby, 1965; Morse, 1964); however, most of these dealt with a very broad, global concept of the self. Self-evaluation as used in this study is strictly concerned with the individual's evaluation of his "presumed goodness for a specific enterprise" (Diggory, 1966, p. 202). In other words the subject does not rate himself on a specific ability, but rather he rates himself as a person who needs that ability to function in that "enterprise."

The summary on self-evaluation research by Diggory (1966) indicates that, on criterion-referenced evaluations, an individual's self-evaluation will drop after failure. For the same reasons presented in

the above discussion on task-related affect, it is assumed here that self-evaluation will drop less for the negotiated criterion-referenced group than for the regulated criterion-referenced group.

Most research on self-evaluation and norm-references has dealt with self-evaluation in a much broader scope than used above. Bloom (1976) presents a summary of research along this line which indicates that, over a period of time, failures under a norm-referenced system lead to a progressive lowering of academic self-concept. Because of the difficulty in seeing objective improvement under norm-references, it is believed that self-evaluation will be lower under norm-references than either regulated or negotiated criterion-references.

From this discussion, the eighth null hypothesis is generated:

8) There are no differences between the three evaluation methods on self-evaluation.

CHAPTER III

METHOD AND PROCEDURE

Subjects

The subjects were selected from sixth-grade students attending two rural public elementary schools in Central Oklahoma. The teachers of the sixth-grade students were asked to designate those students whose school performance was in the average range based on overall grade average and teacher evaluation. The "average range" was defined as between a low "B" and low "C" grade. Of this population, eighteen males and eighteen females were randomly selected and randomly assigned to three groups of twelve students, six males and six females to each group. The three groups were assigned to one of the three evaluation methods: norm-referenced, regulated criterion-referenced, and negotiated criterion-referenced.

Instruments

Seventh Grade Readiness Test

This "test" is actually a matching task comprised of forty three-by-five inch cards each printed with one of eight geometric designs, and a fourteen-by-twenty-two inch mat displaying each of the eight designs. (See Figure 1.) Also included was a small notebook ("test manual") containing instructions and norms. The object of the test is

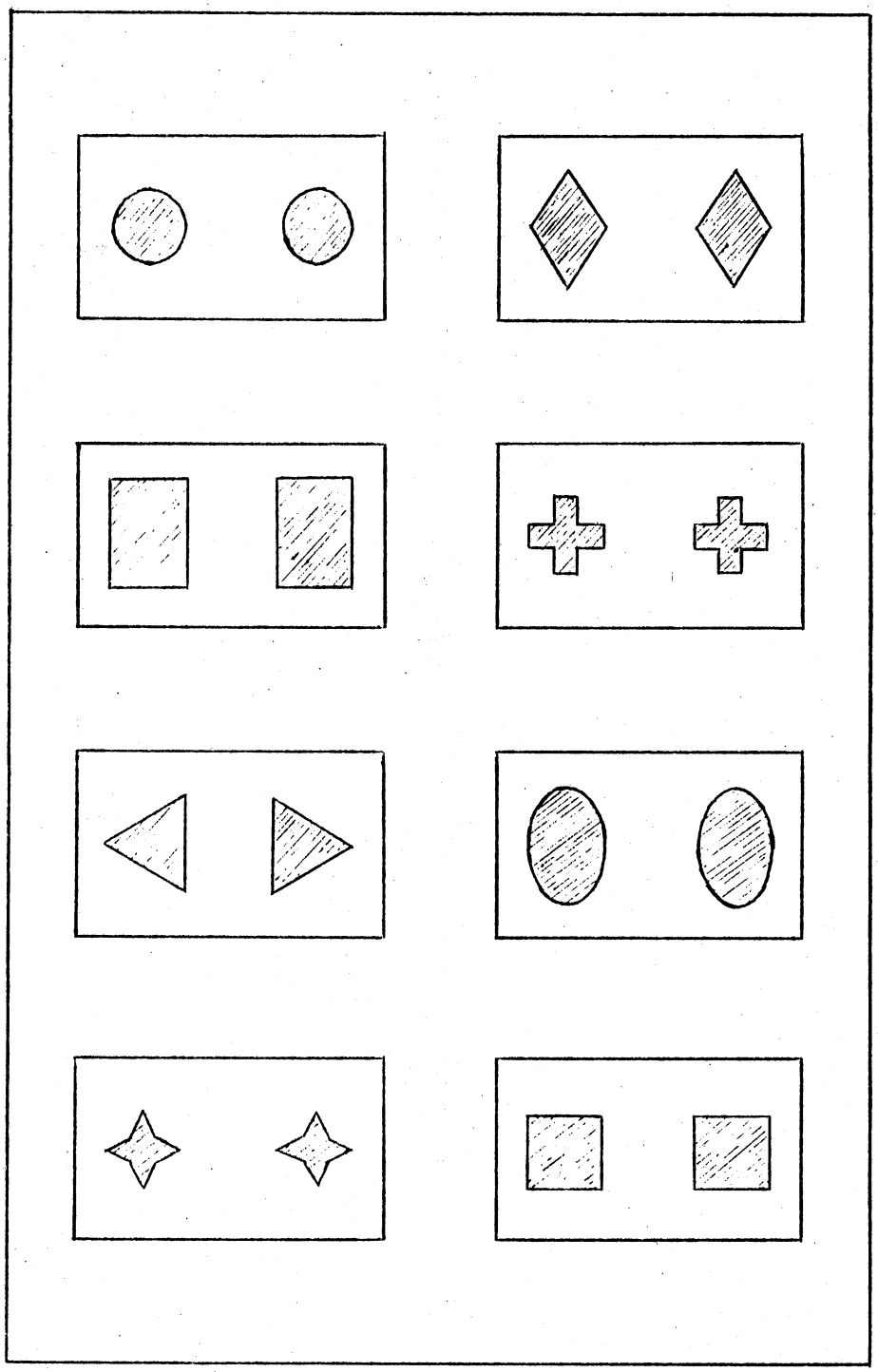


Figure 1. Geometric Designs Displayed on the Testing Mat

to sort as many cards as possible into the eight categories within a twenty second time limit. Each subject had ten trials. Although the test does not measure "seventh-grade readiness," the subjects were lead to believe that it could do so.

The following discussion on the four response forms will describe the forms, and present some reliability and validity information. When reading the information presented here, it must be kept in mind that the following are one-item instruments which typically have lower reliabilities than multi-item instruments. The coefficients presented are measures of stability (test-retest); however, they are contaminated with probable differential effects since the experiment occurred between response forms. If one were to assume that each member of an experimental group was affected equally by the evaluation method, then the coefficients should be near 1.00. When working with psychological variables, however, such an assumption is generally unfounded. It might be more practical to view the test-retest correlations as reflecting differential learning effects more than stability. The coefficients reported are in terms of Spearman's rho which was modified for use with many tie scores (Edgington, 1969). This modified formula for many ties tends to lower values of rho (compared to the usual formula), but reflects a more accurate correlation. Rho coefficients for task-related affect and self-evaluation were computed on pre- and post-measurement scores for those variables. Level of aspiration and probability of success variables were assessed before each trial and rho coefficients were computed between the first and fourth, first and seventh, and first and tenth trials to reflect correlational trends.

The elapse time between the pre- and postmeasurement of task-related affect and self-evaluation was approximately thirty minutes. The elapse time between the paired trials of level of aspiration and probability of success were ten, twenty, and thirty minutes. A validity inquiry was conducted by questioning the subject after the experiment as to what each response form was asking.

Self-Evaluation Scale

A self-evaluation scale (pre- & post-) was used in which the subject estimated his "goodness" in terms of a specific ability on a linear scale divided for each of the following adjectives: poor, fair, good, very good, superior. The subject was to rate himself according to the question, "How would you rate yourself as a potentially successful seventh grade student?" (See Figure 2.)

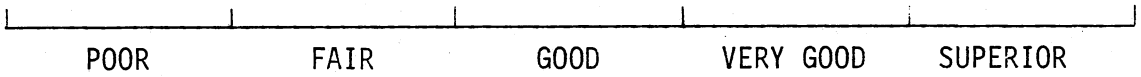
Rho coefficients for the norm-referenced, regulated criterion-referenced, and negotiated criterion-referenced evaluation methods on self-evaluation were .38, .09, and .25 respectively. The subjects described the question by stating something similar to, "It wants to know if I think I will do good or bad in the seventh grade."

Task-Related Affect Scale

Task-related affect was measured on a linear scale divided for each of the following: unwilling, barely willing, moderately willing, very willing, definitely willing. The subject rated himself before and after the test to the following question: "How willing would you be to take similar tests in the near future?" (See Figure 2.)

SEVENTH GRADE READINESS TEST

How would you rate yourself as a potentially successful seventh grade student?



How willing would you be to take similar tests in the near future?

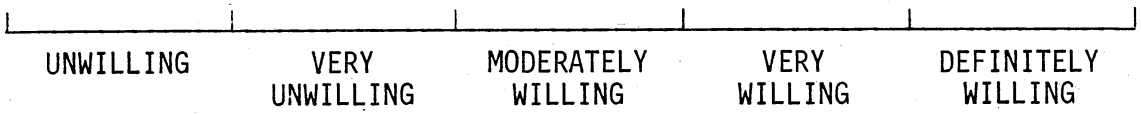


Figure 2. Self-Evaluation and Task-Related Affect Response Form

Rho coefficients for task-related affect were observed as follows: norm-referenced .77, regulated criterion-referenced .87 and negotiated criterion-referenced .58. In response to the inquiry of what was being asked, the subjects typically used words such as "like," "enjoy" and "fun" to describe willingness to take similar tests.

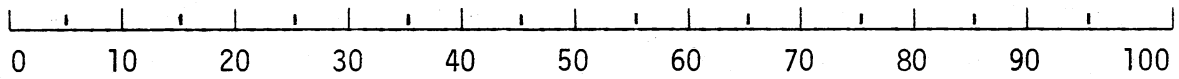
Probability of Success Scale

The subject's attitude was also measured by his estimation of his "power" to achieve a goal: his probability of success. Preceding each trial the subject estimated his probability of succeeding within the remaining trials on a linear scale marked from 0% to 100%. For the regulated criterion-referenced groups, the subjects responded to the question, "What do you think are your chances of being classified as a potentially successful seventh-grade student by scoring 30 or more on this test?" (See Figure 3.) For the negotiated criterion-referenced group the stimulus was, "What do you think are your chances of being classified as a potentially successful seventh-grade student by scoring _____ or more on this test?" (See Figure 4.) The blank was filled in with the negotiated criterion number. The subjects of the norm-referenced group responded to the question, "What do you think are your chances of being classified as a potentially successful seventh-grade student by scoring at the 75th percentile or higher on this test?" (See Figure 5.)

The rho coefficients for probability of success were formulated between trials one and four, one and seven, and one and ten of each evaluation method. They are presented in the above order respectively:

SEVENTH GRADE READINESS TEST

What do you think are your chances of being classified as a potentially successful seventh grade student by scoring 30 or more on this test?

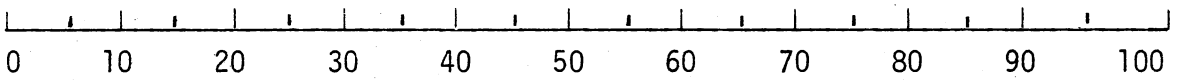


What score are you going to try to make on the next trial? _____

Figure 3. Probability of Success and Level of Aspiration Response Form for Regulated Criterion-Referenced Evaluation Method

SEVENTH GRADE READINESS TEST

What do you think are your chances of being classified as a potentially successful seventh grade student by scoring ____ or more on this test?

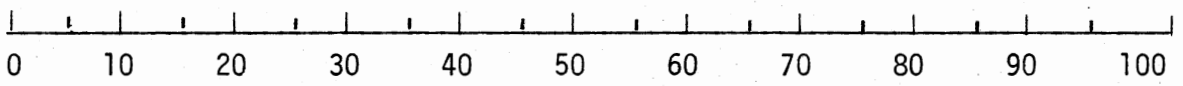


What score are you going to try to make on the next trial? _____

Figure 4. Probability of Success and Level of Aspiration Response Form for Negotiated Criterion-Referenced Evaluation Method

SEVENTH GRADE READINESS TEST

What do you think are your chances of being classified as a potentially successful seventh grade student by scoring at the 75th percentile or higher on this test?



What score are you going to try to make on the next trial? _____

Figure 5. Probability of Success and Level of Aspiration Rresponse Form for Norm-Referenced Evaluation Method

norm-referenced .62, .29, .65; regulated criterion-referenced .66, .47, .65; negotiated criterion-referenced .81, .87, .77. The subjects used phrases such as "my ability to do it," "how much I think I can," and "if I'm sure or not."

Level of Aspiration Form

The subject's immediate goal for the next trial (level of aspiration) was assessed preceding each trial. The subject wrote down a number in response to the question, "What score are you going to try to make on the next trial?" (See Figures 3, 4, and 5.)

The rho coefficients for level of aspiration were also paired as with probability of success. The coefficients are presented in a one to four, one to seven, and one to ten order for each group: norm-referenced .39, .02, .05; regulated criterion-referenced .49, .32, -.27; negotiated criterion-referenced .00, .00, .06. Subjects typically stated that this question wanted to know what the next score was that they were going to try to make.

Variables

Independent: evaluation method.

1) Norm-referenced failure.
2) Regulated criterion-referenced failure.
3) Negotiated criterion-referenced failure.

Dependent: attitudinal measures.

1) Self evaluation was assessed with a linear scale from 1 to 5.
2) Probability of success was assessed with a linear scale from 0% to 100%.

3) Level of aspiration was assessed by fill-in-the-blank from 0 to 40.

4) Task-related affect was assessed with a linear scale from 0 to 5.

Procedure

The procedure of the actual research has six general steps which are repeated for each subject: 1) review the "test" purpose, 2) explain the rationale of the "test", 3) explain how to do the "test", 4) explain the evaluation method, 5) perform the 10 trials, and 6) de-brief the subject. All testing was done individually lasting approximately thirty minutes.

For the first three steps the subject was brought to the examination room and asked if anyone had told them about the test. This was done to insure that only naive subjects participated. At this point the subject was asked to rate himself as a potentially successful seventh-grade student. The subject was told to cover his response so that only he knew how he rated himself. This was stressed with all scales and forms so as to decrease the likelihood of socially acceptable responses. The subject was asked to rate himself before knowing anything about the "test" so that a rating uncontaminated by the information could be obtained. The following was then stated by the examiner:

> The school principal is starting a new procedure this year for sixth-grade students to help decide which students will be successful seventh-graders and which students will have trouble in the seventh-grade. I am going to give you a test which will tell us if you are going to be a good or poor seventh-grader.
>
> The following was read from the "test manual":
>
> This is a card sorting test which consists of forty cards and on each card there is pictured one of eight geometric designs. We have learned that successful seventh-grade students can quickly and correctly sort these cards into

> groups that go together. What you are to do is sort as many of these cards as possible into their groups in twenty seconds. You will have ten chances to do this before I stop the testing. Do you have any questions?

At this point the subject was required to rate his willingness to participate in similar tests at a later time. This form, and all others, were placed upside down in a box, after the subject responded, in a further attempt to remove external pressure in responding.

Step four, explaining the evaluation method, differed depending on which group the subject belonged, and the instructions read differed accordingly. As the subject was read the instructions, his attention was called to a progress sheet used to record his performance. (See Figures 6 and 7.) This allowed the subject to monitor his progress visually in addition to the auditory feedback of the examiner. The instructions read were as follows:

> Regulated criterion-referenced evaluation:
>
> Your performance will be evaluated in terms of whether you reach the criterion score of 30 correctly sorted pictures in twenty seconds on any one trial. This means that you must get a score of 30 to be a successful seventh-grader.
>
> Negotiated criterion-referenced evaluation:
>
> The successful seventh-grader is one who can set realistic goals for a task and meet those goals. You are to evaluate your own card sorting ability with the help of the examiner and set a goal before starting the test. This means that you are to decide how many cards you must sort in twenty seconds to be a successful seventh-grader. The score should be high enough so that it is not easily reached without some effort.

The examiner then guided the subject in setting a goal of that required of the regulated criterion-referenced group. This was accomplished by stating that trying to make a score of forty meant sorting two cards a second, which would be difficult, and that trying

SEVENTH GRADE READINESS TEST

NAME: _____ SCHOOL: _____ DATE: _____

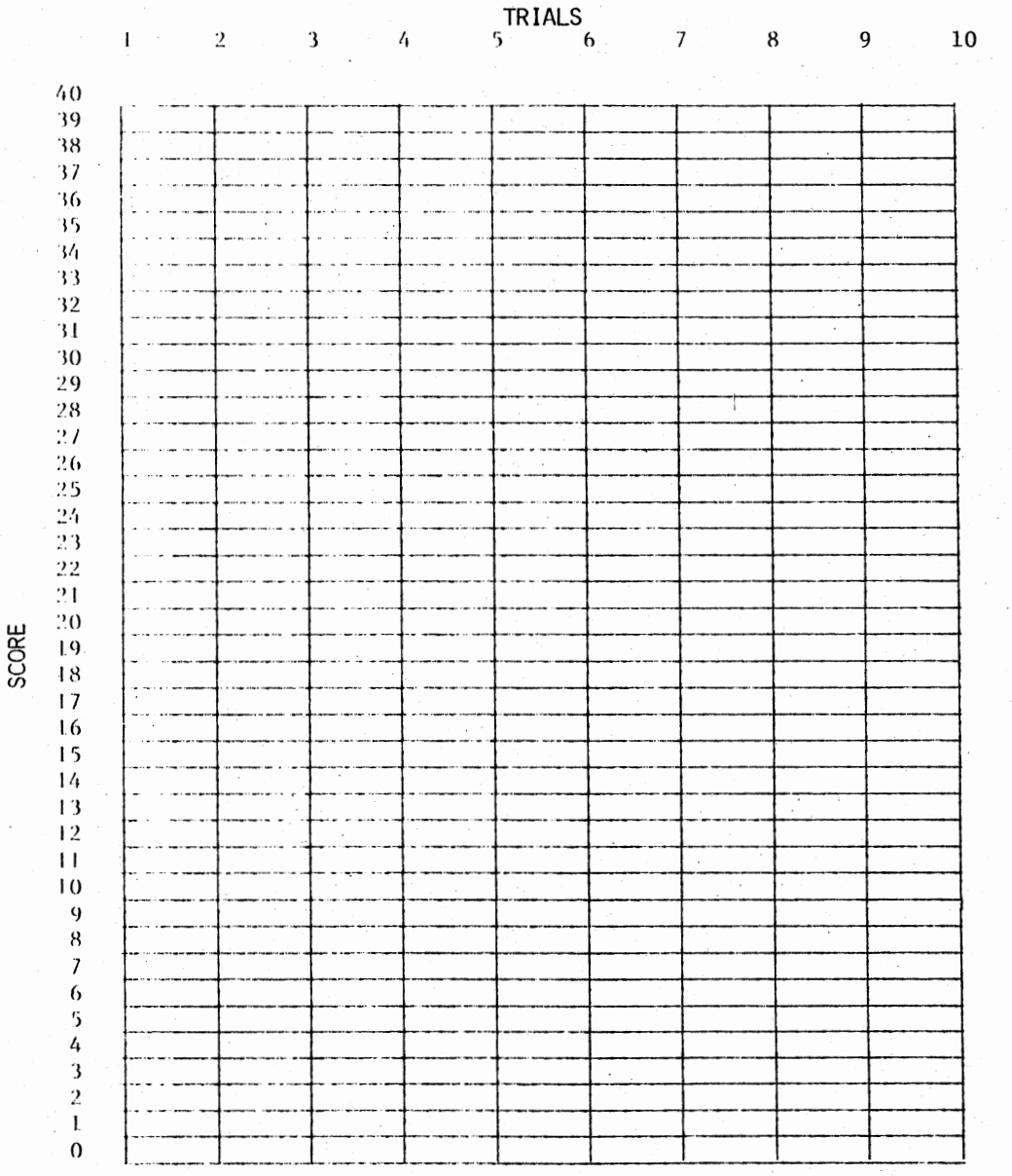


Figure 6. Progress Sheet for Regulated and Negotiated Criterion-Referenced Evaluation Methods

SEVENTH GRADE READINESS TEST

NAME:_____ SCHOOL:_____ DATE:_____

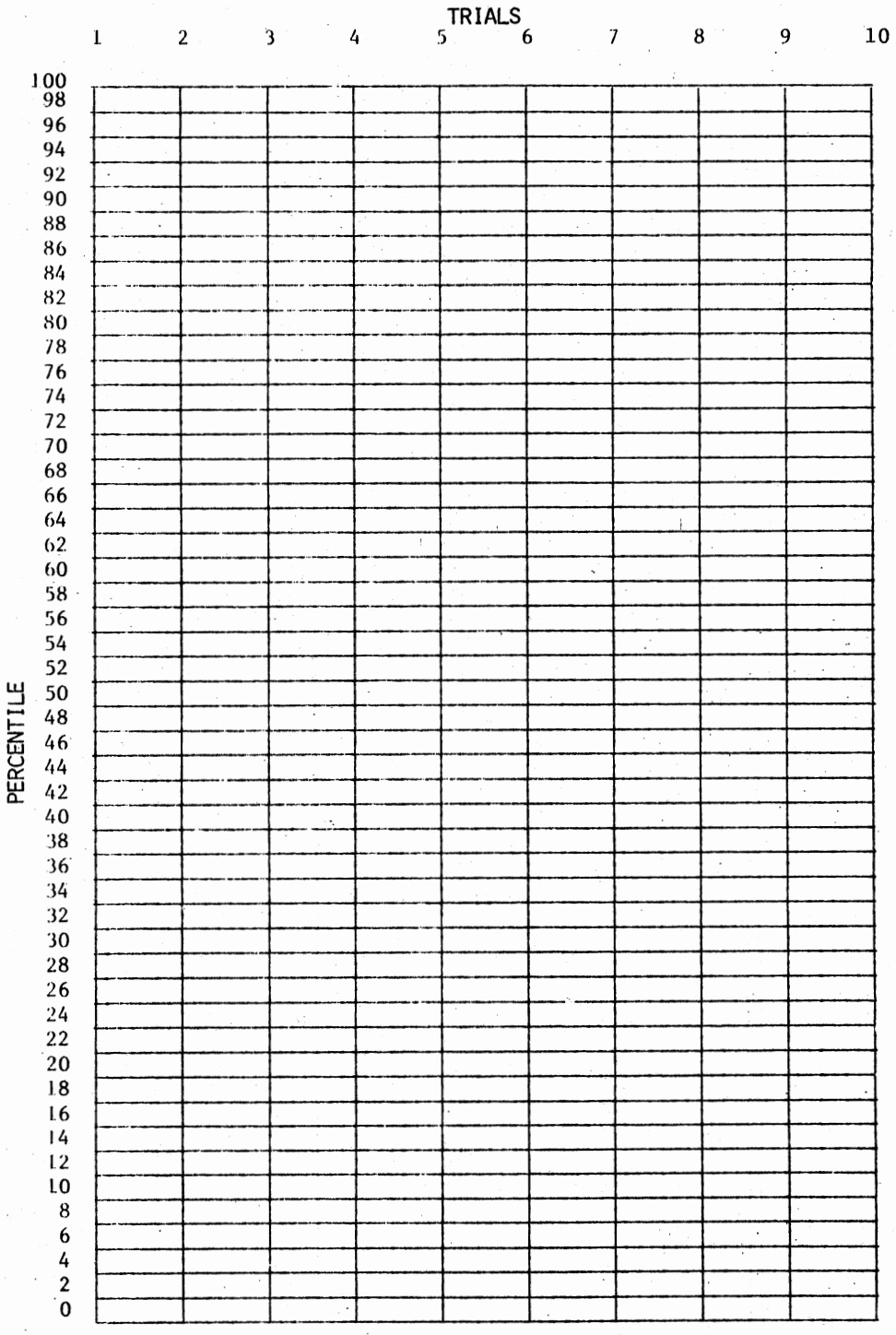


Figure 7. Progress Sheet for Norm-Referenced Evaluation Method

to achieve a score of twenty meant one card each second, a relatively easy task. The subject was then asked what score he wanted to strive for and the examiner encouraged an increase or decrease until thirty was agreed upon. Many subjects selected thirty initially which required the examiner to thoughtfully agree with their decision. The examiner found it to be a relatively easy task to "guide" those subjects initially selecting a score other than thirty to agree upon a score of thirty.

For the norm-referenced group the instructions were:

> Your performance will be evaluated in terms of how well you score compared to how well other sixth-graders score. A sixth-grade student who will be a successful seventh-grader is one who can score at or above the 75th percentile on any one trial. This means that you must score better than 3/4 of your sixth-grade friends.

These instructions were elaborated on by pointing to the progress sheet and stating that a percentile of 100 meant all other sixth-graders did poorer on the test, a percentile of 0 meant all other sixth-graders did better on the test, while a percentile of 50 meant half did better, half did poorer.

On each of the progress sheets the examiner underlined either the score 30 or the percentile 75 depending upon the group. Doing this allowed the subject to have a visual reference point and the progress sheet remained in full view throughout the testing.

The fifth step is that of actually performing to 10 trials. This includes having the subject estimate his probability of success and level of aspiration on the first trial and to do so before each of the remaining 9 trials. The probability of success estimates were in terms of reaching a specific criterion for the two criterion-referenced

groups, while it was in terms of being in the upper quarter of the norm group for the norm-referenced group.

Each subject began the series of 10 trials believing they had twenty seconds for each trial. Actually, the examiner manipulated the reporting of the elapsed time. This was done by predetermining the scores which each subject would achieve, and simply saying "time's up" when the subject reached that score for that trial. Many subjects showed surprise with how quickly "twenty" seconds passed to which the examiner typically replied, "It's easy to lose track of time when you're so involved in the test, isn't it?" All subjects agreed it was and continued with the remaining trials without again questioning the time.

The scores and rate of increase of the scores were reported to the subjects as follows:

Trial	1	2	3	4	5	6	7	8	9	10
Score	9	17	20	22	23	24	24	25	25	25

This negatively accelerating performance curve is graphically displayed in Figure 8.

After each trial, the examiner would inform the subject of the achieved score, but in the norm-reference group the examiner converted this into the subject's relative standing (percentile rank) in the group which were reported to these subjects as follows:

Trial	1	2	3	4	5	6	7	8	9	10
Percentile	45	49	50	48	49	53	50	52	51	50

These percentile ranks are illustrated in Figure 9.

Upon completion of the 10 trials, the subject filled out another self-evaluation rating scale and task-related affect was also assessed by completing that scale.

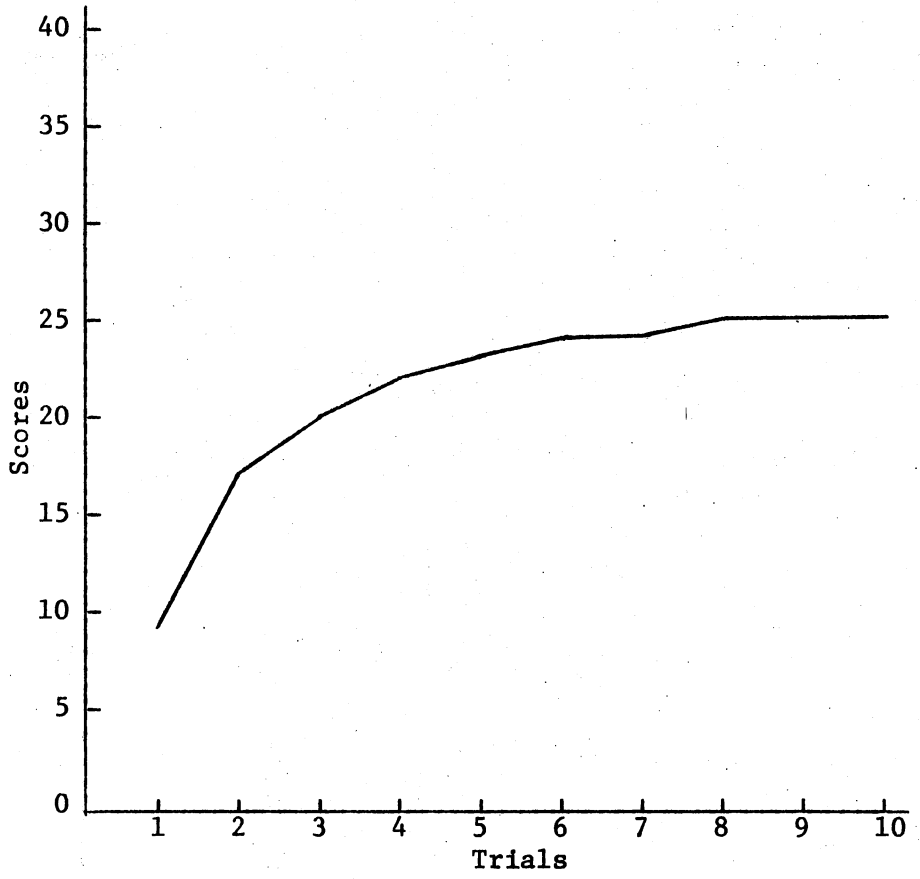


Figure 8. Scores Reported to all Subjects by Trials: the Performance Curve

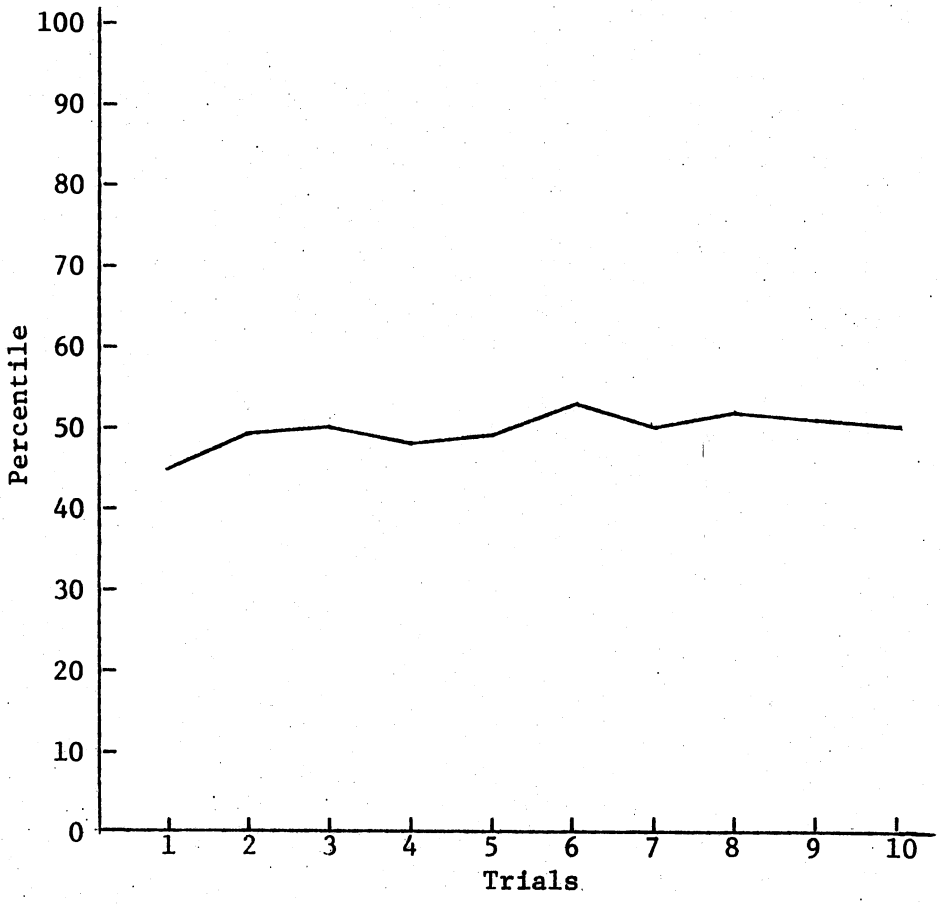


Figure 9. Percentiles Reported to Subjects of the Norm-Referenced Evaluation Method by Trials

The sixth step involves briefing the subject before he leaves so that he is aware that this "test" was not a real evaluation of potential successfulness. To do this, the examiner showed the subject his list of predetermined scores or percentiles and explained how these were the only possible "scores." Most subjects breathed a sigh of relief and stated they were glad that it did not measure successfulness, which attests to the seriousness they prescribed to the testing. The subject was given a brief explanation of the real purpose of the study and the examiner secured their cooperation in maintaining a silence as to the purpose.

Hypotheses

The following eight hypotheses are stated in the null:

1) There are no differences between the three evaluation methods on level of task-related affect.

2) There are no differences between the ten trials on level of aspiration.

3) There are no differences between the three evaluation methods on level of aspiration.

4) There are no interactions between evaluation methods and trials on level of aspiration.

5) There are no differences between the ten trials of probability of success.

6) There are no differences between the three evaluation methods on estimates of the probability of success.

7) There are no interactions between evaluation methods and trials on probability of success.

8) There are no differences between the three evaluation methods of self-evaluation.

Statistical Analysis

The research design is a 3 x 10 factorial with 3 representing the evaluation methods and 10 being the repeated trials. The data for level of aspiration and probability of success was analyzed by means of a split plot analysis of variance (SPF 3.10) with repeated measures on the ten level factor (Kirk, 1968). The data for self-evaluation and task-related affect was analyzed by means of one-way analysis of variance. All a posteriori comparisons among means were made with Tukey's ratio (Hopkins & Glass, 1978; Kirk, 1968) and a minimum significance level of .05 was selected for F and q ratios.

CHAPTER IV

RESULTS

Introduction

The purpose of this study was to examine the attitudinal differences resulting from three evaluation methods after failure on a task. The attitudinal variables measured were: task-related affect, level of aspiration, probability of success, and self-evaluation. The three evaluation methods were: norm-referenced, regulated criterion-referenced, and negotiated criterion-referenced. Analysis of variance with repeated measures was performed on level of aspiration and probability of success. One way analysis of variance was used with task-related affect and self evaluation. Tukey's ratio was performed on significant differences, and a minimum of .05 was selected for significance.

Tests of the Hypotheses

The eight null hypotheses will each be discussed in terms of the statistical results obtained from one-way analysis of variance, split plot analysis of variance with repeated measures, and Tukey's ratio.

The first null hypothesis stated: There are no differences between the three evaluation methods on level of task-related affect. The summary of the one way analysis of variance for the three evaluation

methods on task-related affect is located in Table I. The results indicate a significant difference between the three evaluation methods

TABLE I

ANALYSIS OF VARIANCE SUMMARY FOR TASK-RELATED AFFECT

Source	SS	df	MS	F	P
Between	9.556	2	4.778	9.461	$< .01$
Within	16.667	33	0.505		
Total	26.223	35			

for task-related affect [$F\ (2,33) = 9.46$, $p < .01$]; therefore, the first null hypothesis was rejected. Tukey's ratio between the three possible pairs of means was performed, and the results indicate a significant difference between the norm-referenced evaluation method and the regulated criterion-referenced method [$q\ (3,33) = 4.87$, $p < .01$]. A significant difference was also found to exist between the norm-referenced and negotiated criterion-referenced evaluation methods [$q\ (3,33) = 5.68$, $p < .01$]. No significant difference occurred between the regulated and negotiated criterion-referenced methods. Figure 10 displays the means of the three evaluation methods for task-related affect.

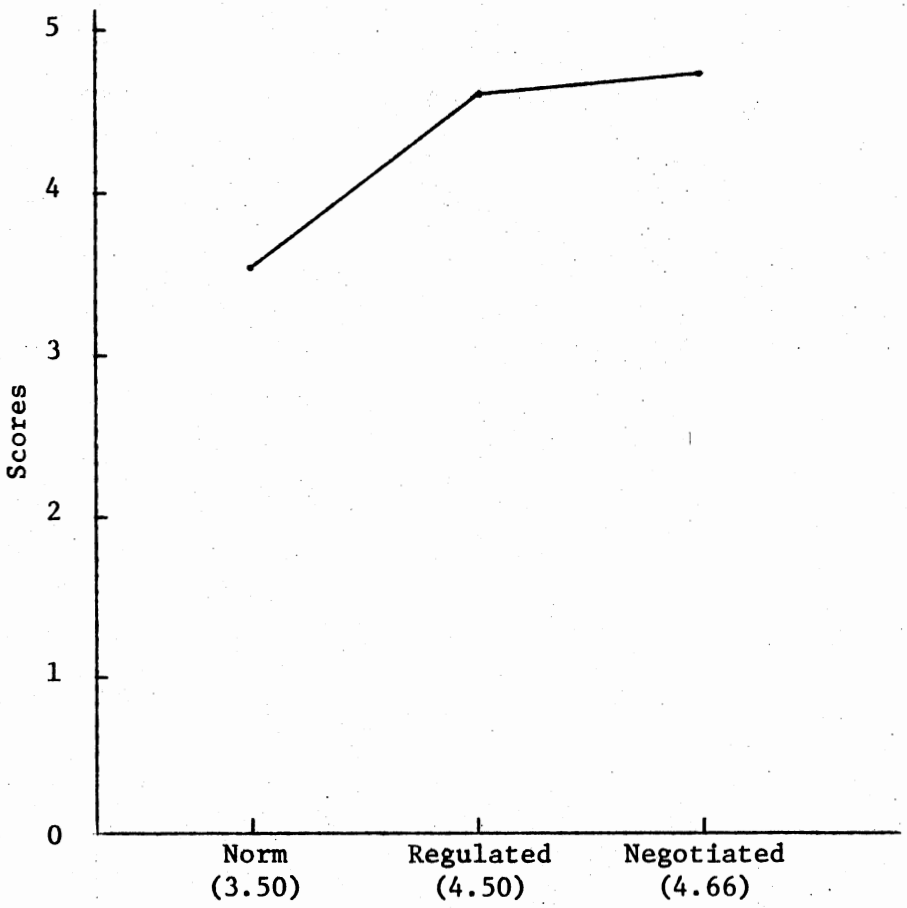


Figure 10. Means by Evaluation Method for Task-Related Affect

The second null hypothesis stated: There are no differences between the ten trials on level of aspiration. Table II presents the summary of the repeated measures analysis of variance for level of aspiration. The results indicate a significant difference between trials for level of aspiration [$F\ (9,297) = 48.61$, $p < .01$], and therefore the rejection of the second null hypothesis. Tukey's ratio

TABLE II

ANALYSIS OF VARIANCE SUMMARY FOR LEVEL OF ASPIRATION

Source	SS	df	MS	F	p
Total	11048.23	359			
Between subjects	3017.63	35			
methods	179.57	2	89.78	1.04	$> .05$
$error_b$	2838.05	33	86.00		
Within subjects	8030.60	324			
trials	4695.11	9	521.67	48.61	$< .01$
trials x methods	148.04	18	8,22	0.76	$> .05$
$error_w$	3187.44	297	10.73		

between adjacent pairs of trial means for level of aspiration revealed three significant combinations: trials 1 and 2 [$q\ (10,297) = 11.55$,

$p < .01$], trials 2 and 3 [q (10,297) = 8.39, $p < .01$], and trials 3 and 4 [q (10,297) = 6.56, $p < .01$]. All other adjacent trials were nonsignificant. The trial means for level of aspiration are located in Table III.

The third null hypothesis reads: There are no differences between the three evaluation methods on level of aspiration. Table II presents the analysis of variance results for level of aspiration. No significant differences were found between the three evaluation methods on level of aspiration; therefore, the third null hypothesis was accepted. The means for the three groups are: norm-referenced = 24.86, regulated criterion-referenced = 26.20, and negotiated criterion-referenced = 24.59.

The fourth null hypothesis stated: There are no interactions between evaluation methods and trials on level of aspiration. Table II reports the analysis of variance for level of aspiration. No significant interaction effect was obtained between methods and trials on level of aspiration (See Figure 11); therefore, the fourth null hypothesis was accepted.

The fifth null hypothesis proposed that: There are no differences between the ten trials on probability of success. Table IV presents the summary of the repeated measures analysis of variance for probability of success, and Table V presents the trial means for probability of success. No significant differences were found between the ten trials of probability of success which lead to the acceptance of the fifth null hypothesis.

TABLE III

MEAN SCORES FOR TRIALS ON LEVEL OF ASPIRATION

Trial	1	2	3	4	5	6	7	8	9	10
Mean	23.44	17.13	21.72	25.30	26.16	26.63	27.11	28.02	28.16	28.77

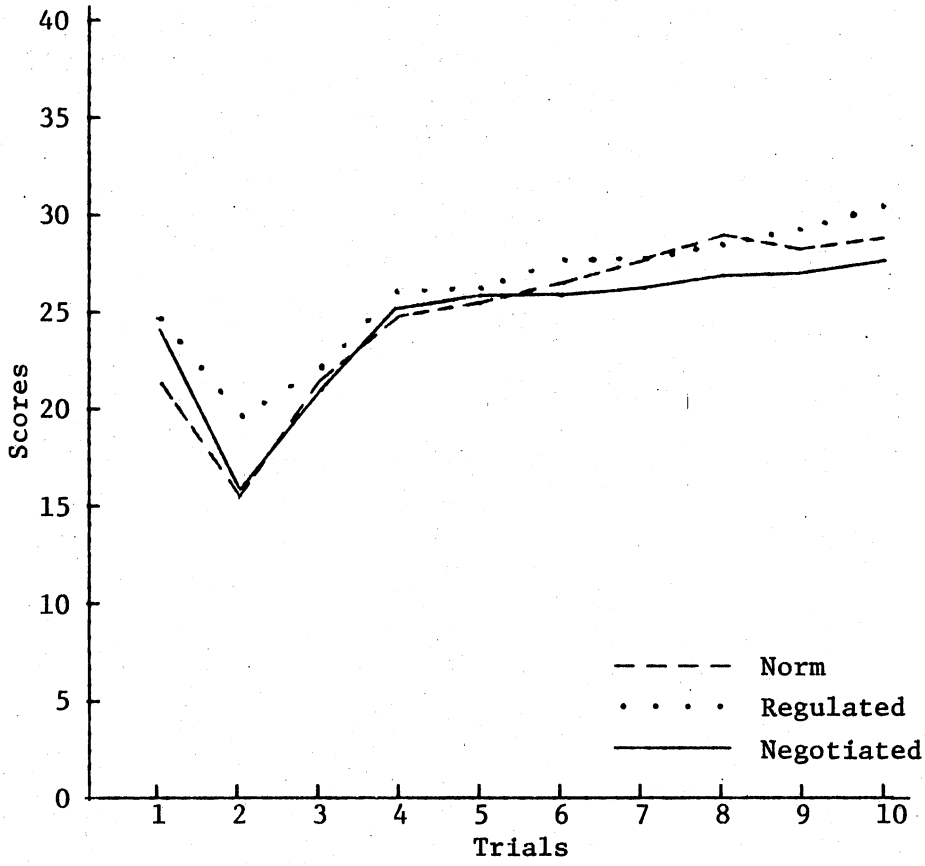


Figure 11. Level of Aspiration by Trials and Evaluation Method

TABLE IV

ANALYSIS OF VARIANCE SUMMARY FOR PROBABILITY OF SUCCESS

Source	SS	df	MS	F	p
Total	1221.00	359			
Between subjects	530.72	35			
methods	127.74	2	63.87	5.23	$<.05$
$error_b$	402.98	33	12.21		
Within subjects	690.27	324			
trials	25.53	9	2.83	1.32	$>.05$
trials x methods	30.07	18	1.67	0.78	$>.05$
$error_w$	634.66	297	2.13		

The sixth null hypothesis reads: There are no differences between the three evaluation methods on estimates of the probability of success. The analysis of variance results for probability of success are located in Table IV. The results indicate a significant difference between the three evaluation methods for probability of success [$F\ (2,33) = 5.23$, $p < .05$]; thus, the sixth null hypothesis was rejected. Tukey's ratio between the three possible pairs of means was performed and the results indicate a significant difference between norm-referenced and regulated criterion-referenced evaluation methods [$q\ (3,33) = 3.80$, $p < .05$]. A significant difference was also found between the norm-referenced

TABLE V

MEAN SCORES FOR TRIALS ON PROBABILITY OF SUCCESS

Trial	1	2	3	4	5	6	7	8	9	10
Mean	53.19	44.30	52.08	51.90	50.69	50.13	50.69	50.41	50.69	45.69

method and the negotiated-criterion method [q (3,33) = 4.11, $p < .05$]. No significant difference was obtained between the regulated and negotiated criterion-referenced evaluation methods. Figure 12 illustrates the means of the three evaluation methods of probability of success.

The seventh null hypotheses stated: There are no interactions between evaluation method and trials on probability of success. The results presented in Table IV indicate no significant interactions; and thus, the acceptance of the seventh null hypothesis. Figure 13 displays the means by trials for the three evaluation methods on probability of success.

The eighth and last null hypothesis proposed that: There are no differences between the three evaluation methods on self-evaluation. Table VI summarizes the analysis of variance results for self-evaluation for which a significant difference was obtained [F (2,33) = 7.98, $p < .01$]. Tukey's ratio was performed between the three possible pairs of means and the results indicate a significant difference between the norm-referenced evaluation method and the regulated criterion-referenced method [q (3,33) = 3.70, $p < .05$]. A significant difference was also obtained between the norm-referenced method and the negotiated criterion-referenced method [q (3.33) = 5.55, $p < .01$]. No significant difference was found between the regulated and negotiated criterion-referenced evaluation methods. Figure 14 illustrates the means of the three evaluation methods for self-evaluation.

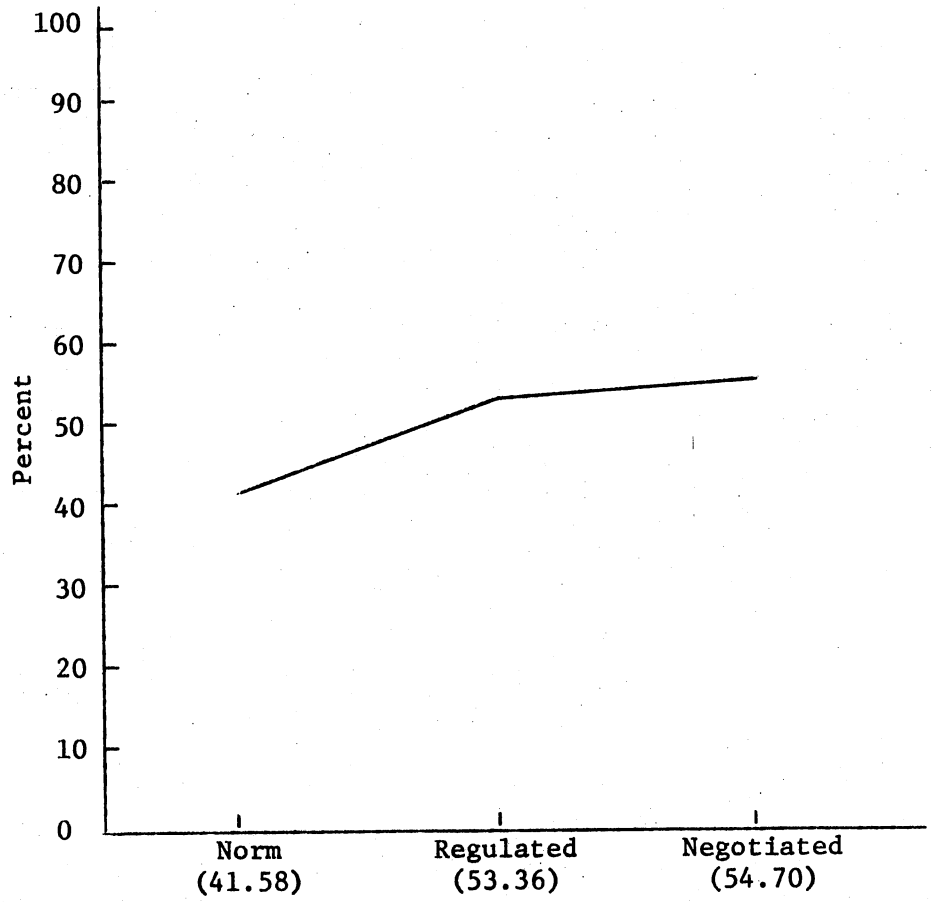


Figure 12. Means by Evaluation Method for Probability of Success

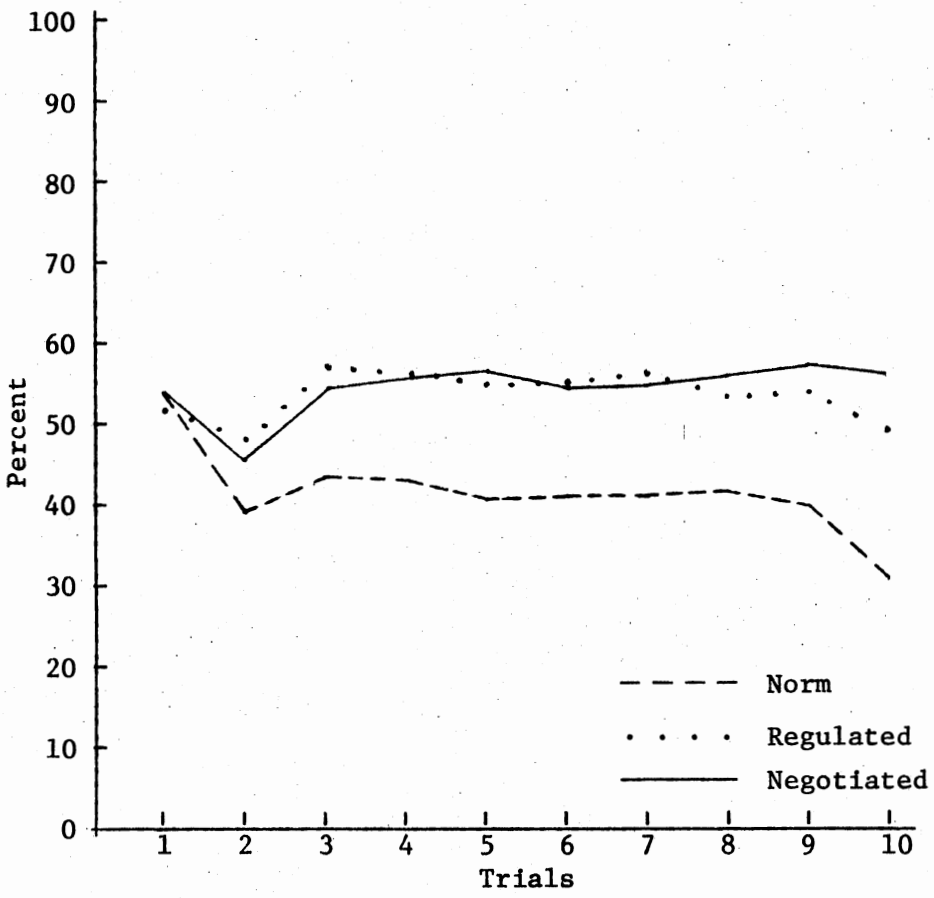


Figure 13. Probability of Success by Trials and Evaluation Method

TABLE VI

ANALYSIS OF VARIANCE SUMMARY FOR SELF-EVALUATION

Source	SS	df	MS	F	p
Between	3.499	2	1.749	7.986	$< .01$
Within	7.251	33	0.219		
Total	10.750	35			

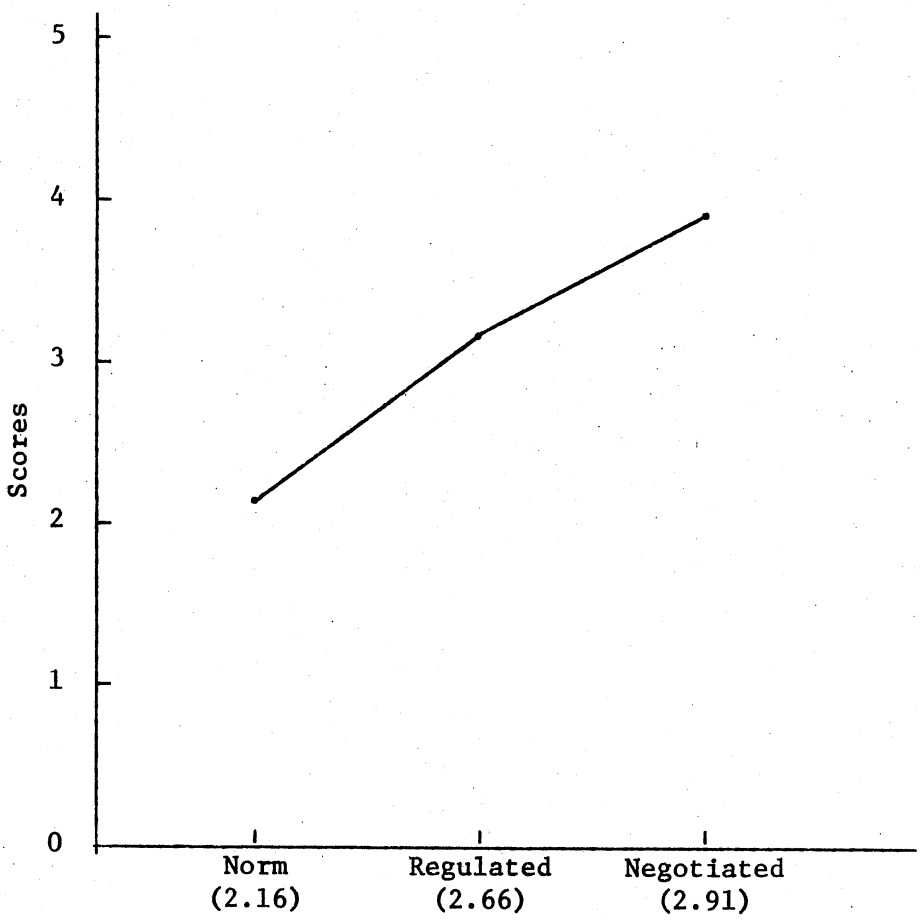


Figure 14. Means by Evaluation Method for Self-Evaluation

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary of the Investigation

This study examined the attitudinal differences resulting from three evaluation methods after failing to succeed on a task. The subject population consisted of sixth grade students selected from two public elementary schools in central Oklahoma. Of this population, eighteen males and eighteen females were randomly selected according to their school performance being in the average range as assessed by overall grade average and teacher evaluation. These thirty-six students were randomly assigned to three groups of twelve students, six males and six females, each to represent the three evaluation methods. The three methods consisted of differences in the way that "success" was defined and evaluated. The norm-referenced method group received evaluation feedback in terms of their relative standing to the norm. Both criterion-referenced groups were evaluated in terms of whether they reached a specific score; the regulated group being told of the "success" score, while the negotiated group established a "success" score with the examiner's help. In other words, the difference between the evaluation methods was that one group's performance was compared to how others did (norm-referenced), while the second group's performance was compared to an already established score (regulated criterion-referenced),

and the third group's performance was compared to a mutually agreed upon score (negotiated criterion-referenced).

The subjects were seen individually and they were told they would be taking a test which predicts successfulness in the seventh grade. This test was a matching task in which cards of eight different geometric designs were sorted into like stacks within a twenty second time limit. They had ten trials in which to reach the appropriate percentile or score marking potential successfulness. All scores and percentiles reported to the subjects were predetermined, and thus all subjects in a group received the same information.

Before each trial, the subject filled out a form asking that a level of aspiration for that trial be set, and to give an estimation of one's probability of success in reaching the "success" mark. After the tenth trial, each subject completed a form in which they rated themselves and their willingness to participate in similar tests.

The results of the study were subjected to one way analysis of variance (self-evaluation and task-related affect) and analysis of variance with repeated measures (level of aspiration and probability of success) with all significant differences being tested with Tukey's ratio.

The results of the investigation indicate that there were no significant interaction effects between the three evaluation methods and the ten trials for either level of aspiration or probability of success. Also, no significant differences were found between trials for probability of success; however significant differences were obtained between trials for level of aspiration. The first three adjacent trial pairs

(1 and 2, 2 and 3, 3 and 4) were all significant, while all other adjacent trials were nonsignificant. No significant differences were obtained between the three evaluation methods for level of aspiration. The three evaluation methods did, however, significantly differ with Respect to probability of success. The results indicate that the norm-referenced method produced significantly lower probability of success estimates than either of the criterion-referenced methods.

Significant differences between the three evaluation methods were obtained for the self-evaluation and task-related affect variables. The norm-referenced method was significantly lower for self-evaluation than either of the criterion-referenced methods. No significant difference occurred between the regulated and negotiated criterion-referenced methods. As for task-related affect, the norm-referenced method produced less willingness to participate in similar tests than either of the criterion-referenced methods. Again, no significant difference was obtained between the regulated and negotiated criterion-referenced methods.

Conclusions of the Study

The lack of interaction between evaluation methods and trials for level of aspiration was as expected since the level of aspiration variable basically follows the shape of the performance curve. Thus, all three possible interactions did not occur since level of aspiration is more dependent on the performance curve than on methods and trials. The subject sets his immediate goal in relation to his previous performance and not on the basis of a combination between how he is evaluated

and how close he is to success. These findings support the research by Diggory (1966). Although level of aspiration was expected to be higher for the norm-referenced method than either criterion-referenced methods, no significant difference was found. Therefore, not only is level of aspiration not affected by an interaction of method and trial, it is also not affected by the method of evaluation. Apparently, the three groups use only their perception of their past performance to establish an immediate goal. The significant differences found between the early adjacent trial pairs for level of aspiration as opposed to later adjacent pairs was partly due to the change in the subject's perception of the task. In other words, the first level of aspiration was formed without any concrete knowledge of performance and therefore was high. After the first trial, level of aspiration become a function of the reported performance curve and the significant differences occurred in early trials because of the sharp negatively accelerated curve. By the fourth trial, the curve's rate of increase began to level off, thus, producing no significant differences between later adjacent trial pairs.

The expected differences between trials for probability of success did not occur. This would indicate that probability of success was not affected by the performance curve itself, that is, the trials. The expected interaction between methods and trials for probability of success also did not occur. It was thought that the criterion-referenced methods and the norm-referenced methods would increase with trials (norm-referenced at a slower rate) and that at about the midpoint in trials, probability of success would drop for the norm-referenced method and not the criterion-referenced methods. This did not occur however,

and the rise and fall in probability of success was similar for the three methods.

The comparisons between the regulated and negotiated criterion-referenced evaluation methods for self-evaluation, task-related affect, and probability of success yielded raw score differences in the predicted direction, but none were significant. Therefore, statistically speaking there were no differences between the two criterion-referenced evaluation methods for self-evaluation, task-related affect, or probability of success. Apparently, the opportunity to participate in the establishment of a "success" score did not make any more difference to the subjects than if an authority dictated the "success" criterion. These results are not in line with the theoretical conceptions of Combs (1966) and Wilhelhm (1967) which would suggest that the opportunity to participate in the evaluation process would produce more positive attitudes than being excluded. The small sample sizes may have affected the outcome.

Significant differences did occur between the norm-referenced method and either of the criterion-referenced methods for self-evaluation, task-related affect, and probability of success. All observed differences were in the predicted direction. These findings add support to the theory and research of Bem (1970) and Bloom (1976) in that the students' perception of their performance was affected by the evaluation method (external cues) and this, in turn, affected the students' attitude toward himself and the task (internal state).

For probability of success, the norm-referenced method produced lower probability of success estimates than did either criterion-

referenced methods. Presumably the norm-referenced subjects estimated their chances as poorer than the criterion-referenced subjects because they were not progressing toward the "success" percentile eventhough their performance scores increased. In other words, for the norm-referenced group, their relative standing remained virtually unchanged, and eventhough they were told their performance was increasing (like the criterion-referenced groups), that knowledge was not enough to offset their perception in light of their percentile ranks. They saw themselves as having less control or power over achievement.

After the ten trials, the norm-referenced group rated themselves poorer than the criterion-referenced subjects on potential successfulness in the seventh grade (self-evaluation). Although all three groups had fialed to meet the standards for success on the test, the norm-referenced group felt worse about themselves than did the criterion-referenced subjects.

A difference was also found in how willing the subjects were to participate in similar tests. The norm-referenced method produced less willingness than either criterion-referenced method. This difference indicates that there was less pleasure associated with the task for the norm-referenced group than the criterion-referenced groups.

A brief conclusion would be that evaluations based on relative standing have a more detrimental effect on students than do evaluations based on absolute standards because the student does not see progress in performance. Absolute standards focus attention on progress and do not cloud the performance feedback with percentiles. Evaluation by relative standing produces greater feelings of hopelessness, greater

devaluation of the self, and less desire to engage the task than evaluation by absolute standards.

Recommendations

This section will be divided into two parts; one discussing the implication of the present study to education, the other suggesting improvements and future research of the study.

The results of this study suggest that if evaluation is to be part of the educational process, then careful thought should be given to the consequences of the evaluation method used. The present study suggests that norm-referenced evaluation has a more detrimental effect on the learner than does criterion-referenced evaluation. It follows, therefore, that norm-referenced evaluation should be used only when criterion-referenced evaluation will not answer the question concerning evaluation purpose. If the purpose is one of requiring the selection of the best student for some award, honor, or scholarship where only a few positions are available, then norm-referenced evaluation appears to be appropriate. The purpose is one of arranging students from low to high (relative standing) on some measure so as to select the highest. Where the evaluation of the level of an individual's performance is important, however, norm-references should be avoided in favor of criterion-referenced evaluations. Both the ongoing (formative) and final (sumative) evaluations should be incorporated into criterion-references. The final grade on any lesson, test, or curriculum subject should be assigned on the basis of reaching the acceptable criterion.

There are several improvements that could be made in the present study and future versions of it. One area of needed improvement is the development and/or use of dependent measuring instruments of greater realiability and validity. The possibility of using more physiological based measures is worth investigating.

Another area of improvement would be to provide the norm-referenced evaluation group with a progress sheet used by the criterion-referenced groups to display changes in performance in addition to the percentile progress sheet used in this study. This would provide the norm-referenced group with visual feedback along with the auditory feedback of raw scores.

The study of different performance curves than the negatively accelerated one used in the present study may yield different results. Of interest would be a steady rising curve and a positively accelerated curve, each of which may produce different attitudinal results.

Some variations of the evaluations might be made such as having the examiner be less directive in the negotiated criterion-referenced method or the study of a criterion-referenced method in which the subject has complete and total say as to the criterion score. The investigation of differences between types of evaluations when the subjects succeed on the task would also be of interest.

A SELECTED BIBLIOGRAPHY

Alvord, D. J., & Glass, L. W. Relationships between academic achievement and self-concept. Science Education, 1974, 58, 175-179.

Astin, A. N., et al. National Norms for Entering College Freshmen 1966. Washington, D.C.: American Council of Education, 1967.

Bem, D. J. Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychological Review, 1967, 74, 183-200.

Bem, D. J. The epistemological status of interpersonal simulations: A reply to Jones, Linder, Keisler, Zanna, and Brehm. Journal of Experimental Social Psychology, 1968, 4, 270-274.

Bem, D. J. Beliefs, Attitudes, & Human Affairs. California: Brooks/Cole Publishing Co., 1970.

Block, J. H. (Ed.). Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, 1971.

Block, J. H. (Ed.). Schools, Society, and Mastery Learning. New York: Holt, Rinehart and Winston, 1974.

Bloom, B. S. Learning for Mastery, Chapter 3 in B. S. Bloom, J. T. Hasting, & G. F. Madaus, Handbook on Formative and Sumative Evaluation of Student Learning. New York: McGraw-Hill Book Co., 1971.

Bloom, B. S. Human Characteristics and School Learning. New York: McGraw-Hill Book Co., 1976.

Brookover, W. B., et al. Self-concept of ability and school achievement. Sociology of Education, 1964, 37, 271-278.

Carroll, J. B. A Model of School Learning. Teachers College Record, 1963, 64, 723-733.

Cartwright, D. The effect of interruption, completion, and failure upon the attractiveness of activities. Journal of Experimental Psychology, 1942, 31, 1-16.

Centi, P. Self-perception of students and motivation. Catholic Education Review, 1965, 63, 307-319

Combs, A. W. Perceiving, Behaving, Becoming. Washington, D. C.: Association for Supervision and Curriculum Development, 1962.

Combs, A. W. A contract method of evaluation, in Simon, S. B. and Bellance, j. a. Degrading the Grading Myths. Washington D. C.: Association for Supervision and Curriculum Development, 1976.

DeCharms, R. From pawns to origins: toward self-motivation, in Lesser, G. S. Psychology and Educational Practices. Glenview, Ill.: Scott, Foresman and Co., 1971.

Diggory, J. C., & Morlock, H. C., Jr. Level of aspiration, or probability of success? Journal of Abnormal and Social Psychology, 1964, 69, 282-289.

Diggory. J. C. Self-evaluation: Concepts and Studies. New York: John Wiley and Sons, Inc., 1966.

Diller, L. Conscious and unconscious self-attitudes after success and failure. Journal of Personality, 1954, 23, 1-12.

Edgington, E. S. Statistical Inference: The Distribution Free Approach. New York: McGraw-Hill Book Co., 1969.

Feather, N. T. Performance at a difficult task in relation to initial expectation of success, test anxiety, and need achievement. Journal of Personality, 1965, 33, 200.

Gewirtz, H. B. Generalization of children's preferences. Journal of Abnormal and Social Psychology, 1959, 58, 111-117.

Gibby, R. G., & Gibby, R. G., Jr. The effects of stress resulting from academic failure. Journal of Clinical Psychology, 1965, 63, 307-319.

Glasser, R., & Nitko, A. J. Measurement in Learning and Instruction, in Thorndike, R. L., (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Hopkins, K. D., & Glass, G. V. Basic Statistics for the Behavioral Sciences. New Jersey: Prentice-Hall Inc., 1978.

Khan, S. B. Affective correlates of academic achievement. Journal of Educational Psychology, 1969, 60, 216-221.

Khan, S. B., & Weiss, J. The teaching of affective responses. In Travers, R. M. W., (Ed.), Second Handbook of Research on Teaching. Chicago: Rand McNally & Co., 1973.

Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Wadsworth Publishing Co., Inc., 1968.

Kurtz, J. J., & Swenson, E. J. Student, parent, and teacher attitude toward student achievement in school. School Review, 1951, 59, 273-279.

Malpass, L. F. Some relationships between students' perceptions of school & their achievement. Journal of Educational Psychology, 1953, 44, 475-482.

Michael, W. B., Baker, D., & Jones, R. A. A note concerning the predictive validities of selected cognitive & non-cognitive measures for freshmen students in a liberal arts college. Educational and Psychological Measurement, 1964, 24, 373-375.

Morse, W. C. Self-concept in the school setting. Childhood Education, 1964, 41, 195-198.

Peckham, P. D., & Roe, M. D. The effects of frequent testing. Journal of Research and Development in Education, 1977, 10, 40-50.

Popham, W. J., & Husek, T. R. Implications of criterion referenced measurement. Journal of Educational Measurements, 1969, 6, 1-9.

Popham, W. J. Criterion-referenced Measurement: An Introduction. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Proger, B. B., & Mann, L. Crtierion-referenced measurement: the world gray versus black and white. Journal of Learning Disabilities, 1973, 6, 18-30.

Russell, I. L. Motivation for school achievement: measurement and validation. Journal of Educational Research, 1969, 62, 263-266.

Terwilliger, J. S. Assigning grades—philosopical issues and practical recommendations. Journal of Research and Development in Education, 1973, 10, 21-39.

Wilhelms, F. T. Evaluation as Feedback and Guide. Washington, D. C.: Association for Supervision and Curriculum Development, 1967.

Zimbardo, P., & Ebbesen, E. B. Influencing Attitudes and Changing Behavior. Massachusetts: Addision-Wesley Publishing Co., 1970.

VITA 2

Bruce Dale Cook

Candidate for the Degree of

Doctor of Education

Thesis: THE EFFECTS OF FAILURE BY THREE EVALUATION METHODS ON THE ATTITUDES OF STUDENTS

Major Field: Educational Psychology

Biographical:

Personal Data: Born at Topeka, Kansas, November 29, 1950, the son of Paul D. and Wanda M. Cook.

Educational: Attended elementary and junior high school in Borger, Texas, and Bartlesville, Oklahoma; graduated from Sooner High School, Bartlesville, Oklahoma, in May, 1969; attended Northeastern Oklahoma A & M College, Miami, Oklahoma, for two years; transferred to Oklahoma State University and received the Bachelor of Science degree in May, 1973, with a major in Psychology; continued graduate work at Oklahoma State University and completed the requirements for the Master of Science degree in July of 1976 with a major in Educational Psychology; completed the requirements for the Doctor of Education degree in July of 1979.

Professional Experience: Conducted private tutorial sessions in remedial reading and mathematics, 1974-1975; graduate research assistant, Oklahoma State University, Applied Behavioral Studies in Education Department, 1974-1975; reading instructor for Predischarge Education Program/Career Advancement Program, Vance Air Force Base, Enid, Oklahoma, 1975; served as consultant to the State University of New York at Stony Brook on an Engineering Concepts Curriculum Project minicourse, 1975; school psychologist intern at the Regional Education Service Center, Stillwater, Oklahoma, 1975-1976; school psychologist intern at Bi-State Mental Health Foundation, Ponca City, Oklahoma, 1976-1977; graduate teaching assistant, Oklahoma State University, Applied Behavioral Studies in Education Department, 1977-1978; employed with the Oklahoma State Department of Health as a psychologist at the Grady County Guidance Center, Chickasha, Oklahoma, 1978-present.