1960

Reliability and Validity of the Thematic Apperception Test Scored by the Discomfort-Relief Quotient

Donald W. Culbertson
Central Washington University

RELIABILITY AND VALIDITY OF THE THEMATIC APPERCEPTION TEST SCORED BY THE DISCOMFORT-RELIEF QUOTIENT

A Thesis

Presented to

the Graduate Faculty

Central Washington College of Education

In Partial Fulfillment

of the Requirements for the Degree

Master of Education

by

Donald W. Culbertson

December, 1960

APPROVED FOR THE GRADUATE FACULTY

Eldon E. Jacobsen, COMMITTEE CHAIRMAN

Dean Stinson

Maurice L. Pettit

ACKNOWLEDGMENTS

Grateful acknowledgment is extended to Dr. Eldon E. Jacobsen for his advice and encouragement in directing the writing of this paper, and to Dr. Maurice L. Pettit and to Dr. Dean Stinson for their helpful suggestions.

Further acknowledgment is extended to Dr. E. E. Samuelson, Dean of Students; Dr. Roy F. Ruebel, Dean of Graduate Studies; Dr. Ralph Gustafson, Director of Student Teaching; and Mr. Enos Underwood, Acting Registrar, for making available the information needed to complete this investigation.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

THE PROBLEM AND DEFINITION OF TERMS

I. INTRODUCTION

An endless problem in the behavioral sciences is the adequate description of present human performance. More difficult still is the adequate prediction of future performance. Counseling and placement are contingent on prediction data. An abundance of research has been published on the relationship between intelligence or academic aptitude tests and school grades, promotion, and graduation records (21:1-72; 26:1-62; 27:1-43).

Using this one variable, intelligence, only moderate success is obtained in predicting a criterion such as school grades. Such additional variables as high school grades and standardized achievement tests have added significantly to prediction. The investigation continues for additional variables to be added to the present combination so as to more efficiently predict scholastic success, teaching success, and other criterion variables of human performance.

Although still in their infancy, personality variables are being increasingly investigated for possible prediction measures by way of self inventories, supervisor ratings, teacher ratings, peer ratings, etc (8:1-54). Projective techniques are characterized by a global approach to the appraisal of personality. Projective tests have been seen as potentials in the description and prediction sphere, but scoring has proven to be problematic. It has been difficult to score them rapidly and reliably.

In 1947, Dollard and Mowrer reported the application of "A Method of Measuring Tension in Written Documents" stemming from the hypothesis that a comparison of alternations in "Tension" in social casework records would reveal the progress being made. The case reviewed was that of the "Cellini" family. Their results seemed to justify their hypothesis (6:7-8).

A second study of measuring changes in personality was reported by Raimy in 1948. This study, "Self Reference in Counseling Interviews," condensed from a more detailed dissertation, purported to show that by measuring changes in self-evaluation in counseling interviews, the progress of counseling could be traced quantatively (22:153-159). Although derived from different theoretical positions and applied to different kinds of material, both methods of analysis have the same goal: measuring changes in personality.

Dollard and Mowrer's "Tension Index" is commonly referred to as the Discomfort-Relief Quotient (DRQ). This index is obtained by having judges divide the typescript of the protocol into "Clause or Thought-Units" according to definite rules and then classify each unit as an expression of "Discomfort," "Relief," or "O" (lacking in tension or relief). The "Tension Index" (DRQ) is obtained by dividing the number of discomfort units by the number of discomfort units plus the relief units for any given portion of a written document (6:7):

$$\frac{\text{Discomfort Units}}{\text{Discomfort Units plus Relief Units}} = \text{DRQ}$$

Other methods of obtaining units such as word units, sentence units, and page units are described in their study, but the thought unit is recommended as it yields the greatest degree of coincidence between scorers.

The Dollard and Mowrer study suggests that the "Tension Index" might very well be applied to the Thematic Apperception Test (TAT). Such an application was the general purpose of this investigation.

II. PROBLEM

The central purpose of this study then, was to investigate the reliability of the DRQ method of scoring the TAT. A secondary purpose was to investigate some of its validities or inter-relationships with certain variables in a teacher-education program. Specifically, the following problems were investigated: (1) What are the scorer reliabilities or coefficients of rater equivalence? (2) What is the extent of relationship between tension, as measured by DRQ, and success in college as measured by Grade Point Average (GPA)? (3) What relationship exists between DRQ and the number of quarters spent at Central Washington College of Education (CWCE)? (4) What is the extent of relationship between DRQ and ratings of practice teaching success? and (5) What relationship exists between DRQ and first year teaching success as measured by principal and supervisor ratings?

III. DEFINITION OF TERMS

Reliability. This broad term, used in the Education and Psychology fields, indicates the extent to which a test is consistent in measuring whatever it does measure: its dependability, stability, and relative freedom from errors of measurement. Reliability is usually estimated by some form of reliability coefficient or by the standard error of measurement. In this study, scorer reliability refers to a correlation between two scorers, scoring the same protocols, revealing Coefficients of scorer Equivalence (11:456).

Validity. This term means the extent to which a test does the job for which it is used. Validity thus defined has different connotations for various kinds of tests and, accordingly, different kinds of validity evidence are appropriate for them. The validity of a personality test (TAT) is the extent to which the test yields an accurate description of an individual's personality traits or personality organizations (Status

Validity). It may be evidenced by agreement between test results and such other types of evaluation as ratings or clinical classifications, but only to the extent that such criteria are themselves valid (11:576).

Discomfort-Relief Quotient (Tension Index). This is a measuring instrument used to measure tension. This measure yields a relatively reliable, graphic picture of tension changes in a social case record, and conceivably also in an autobiography, psychoanalysis history, or other personal documents (6:1).

Personality. Personality is referred to as something which is unique, persistent, dynamic, social, and organized. It has been pointed out that while it is possible to know about this "something" only by observing an individual's motive patterns, it cannot be said that personality is the individual's motive patterns. It is, rather, that which "holds together" all motive patterns--that which determines all behavior, both attitudinal and expressive. For present purposes, therefore, it shall mean the individual's organization of predispositions to behavior, including predispositions to both directive and expressive behavior (11:332).

Tension. Tension can be defined as a mental strain; stress of mind or feeling; state of exertion or effort; or a strongly excited condition (11:546). Tension in this study has reference to the above mentioned qualities purportedly revealed through the Tension Index (DRQ) applied to TAT protocols.

Thematic Apperception Test (TAT). This is a projective test in which a person is asked to tell a story suggested by each of nineteen pictures and a blank card. The pictures are sufficiently vague to leave much to the imagination of the testee. The test assumes that themes apperceived by the testee in the pictured behavior are those which are important in his or her own life (11:551).

Learning. Learning is a highly general term for the relatively enduring change, in response to a task-demand, induced directly by experience, or the process or processes whereby such changes are brought about. It also denotes becoming able to respond to a task-demand or an environmental pressure in a different way as a result of earlier response to the same task (practice) or as a result of other intervening relevant experiences. The sign of learning is not a shift of response of performance as a consequence of change in stimulus-situations or in motivation, but rather a shift in performance when the stimulus-situation and the motivation are essentially the same (11:289).

Grade Point Average (GPA). This academic rating using the Point Hour Ratio is a weighted index of the grades or marks received for academic work in an American High School or College. Each grade receives a numerical equivalent: e.g., A=4, B=3, etc. Each number is multiplied by the hours of credit assigned the course in which the grade is gained. The resulting products are summed and divided by the total number of credit hours (11:395).

First Year Teaching Evaluation. This is an Administrator's Progress Report of Beginning Teachers. The Central Washington College of Education (CWCE) sends college representatives to each of the school districts where first year teachers are employed. These representatives discuss the progress of each teacher with his immediate school district supervisor (principal or superintendent) who signs the report when completed. In turn this form is returned to the Dean of Graduate Studies and placed in the teacher's permanent file. This is the last official evaluation made by the college on students who have received teaching degrees.

Student Teaching Ratings at CWCE. In 1957 the Supervisor of Student Teachers, along with other members of the Education Department, designed a new method of evaluating students involved in directed teaching. This form, a joint venture between the College Supervisor, the Classroom Teacher, and the Student Teacher, covers such broad areas as Relationship with Children, Effectiveness in Developing Learning Experiences, Teaching Techniques, Professional Outlook and Self Adjustment, and Personal Qualities. The summary rating of these qualities, at the bottom of the scale, was used as the criterion of student teaching success.

Correlation Coefficient (r). The Pearson Product Moment (r) is the most commonly used measure of relationship between paired facts, the tendency of two or more variables or attributes to go hand-in-hand. It ranges in value from minus 1.00 for perfect negative relationship through 0.00 for none or pure chance to plus 1.00 for perfect positive relationship. Eta, or correlation ratio, used in one instance in this study, varies only from 0.00 to 1.00 (11:124).

Variable. This is a quantity which may increase or decrease, continuously or discontinuously, without essential change in that which is or has the quantity. e.g., the rapidity with which a person can react to a stimulus (11:578). It means any identifiable trait or quality which can be measured.

CHAPTER II

REVIEW OF THE LITERATURE

Research on description of the multiple dimensions of human personality and the use of quantitative descriptions in predicting school and teaching success is extensive. Some background is needed regarding (1) prediction of school and teaching success, (2) description of personality via projective tests, and (3) specific application of methods of scoring for a Tension Index as a personality factor.

An abundance of research has been completed in the field of predicting teaching success. Generally speaking, the results indicate that, as yet, no one method has been devised that will predict teaching success to any significant degree. These studies have used many bases for their correlation studies, with only limited success. Most of the research has been written on the relationship between intelligence or aptitude tests and school grades, achievement tests, freshman entrance examinations, administrators' evaluations, etc.

I. PREDICTION OF SCHOOL AND TEACHING SUCCESS

Van Zee's "Study of Professional Course Grades, College Activities, ACE Scores, and High School Grades as Related to Success in First-Year Teaching" found the correlation between high school grades and supervisor's ratings of beginning teachers not significant. He found that the American Council on Education Psychological Examination (ACE) scores when correlated with supervisors' ratings had an r of —.005 for freshman students and +.281 for transfer students (26:28).

One is led to suspect that the numerical descriptions of behavior used as predictors may not be so much at fault as the criterion we are trying to predict.

Hertz, in his doctoral dissertation on "The Relationship Between the Teaching Success of First-Year Elementary Teachers and Their Undergraduate Academic Preparation," suggests, as have so many other studies, that there is no such thing as selecting students or predicting the success of students from any single variable. He contends that a combination of variables would appear to be much more accurate than any single variable for predicting teaching success (14:23).

An investigation of "The Relationship Between Freshman Entrance Examination Scores and Academic Success in the Curriculum of Central Washington College of Education," by O'Donahue, shows that a positive correlation of moderate magnitude exists between academic success in the curriculum of CWCE and academic aptitude as measured by the American Council on Education Psychological Examination. A similar relationship exists between academic suocess and reading ability measured by the Nelson-Denny Reading Test (21:48-49). However, when college grades are used to indicate teaching success, little relationship is shown (26:31).

Barr, one of the most experienced workers in the area of predicting and appraising teaching success, wrote, ". . .it is apparent that the identification and definition of teaching competencies is as yet by no means satisfactory. We do not yet have an adequate definition of teaching efficiency" (4:1448).

Until recently, personality apart from the ability and aptitude components, was not investigated as a potential in this sphere of predicting success in schools and colleges. In the past few years, increased investigation has been carried on in this broad area, as the researchers feel that although it may not be the total answer, it is an all important factor in the problem of predicting success.

Duncan investigated the "Interrelationship Among Certain Personality Tests and Ratings and Their Relationship to Academic Success in a Teachers College." The results indicate some relationship between certain non-academic measures and grades and no relationship between other personality variables and grades. Moderate correlation between a Tension Index from the TAT and dormitory supervisors' ratings on the "Social" section of the Haggerty-Olsen-Wichman Behavior Rating Schedule B (HOW) and the Minnesota Multiphasic Personality

Inventory (MMPI) Social Introversial Scale (Si) suggests that these instruments are measuring similar traits. The relationship between the MMPI Si and MMPI Responsibility Scale (Re) indicates little relationship between the two scales, suggesting that they are not measuring the same things (8: 40-45).

Although correlations between non-academic variables and grades were low, some were statistically significant and suggest that use of such variables in multiple regression equations may well raise grade prediction formulas.

II. PROJECTIVE PERSONALITY TESTS USED FOR PREDICTION

Sentence Completion. A CWCE research paper by DuFrense, dealing with the "Investigation of the Open-End Sentence Concerning Its Use in School Guidance Programs," reports two factors related to the current research. First, he showed that a Tension Index (DRQ) could be used to score a projective sentence completion test with moderately high reliability. Second, he discovered that DRQ scores did not correlate significantly with grades nor with honor points (7:13-15).

Thematic Apperception Test. _(TAT)_. The use of projective tests such as the TAT to sample personality characteristics has been advocated by many experimenters in recent years. The TAT, according to its author, Murray:

> . . .is a method of revealing to the trained interpreter some of the dominant drives emotions, sentiments, complexes, and conflicts of personality. Special value resides in its power to expose the underlying inhibited tendencies which the subject or patient is now willing to admit, or can not admit because he is unconscious of them (20:1).

It is expected that TAT pictures will serve as a sort of screen upon which the subject projects his characteristics, ideas, attitudes, aspirations, fears, and worries in his effort to make up a story to fit the picture (1:598-605).

Test material (nineteen ambiguous pictures and one blank card) have been found to effectively stimulate the imagination, force the subject to deal with certain human situations in his own way, and give the administrator of the test the advantage of using a prototype instrument. The pictures use two psychological tendencies, according to Murray. One, ". . .they draw on their experiences and express their sentiments and needs consciously and unconsciously," and two, ". . .people interpret an ambiguous situation in conformity with their past experiences and present wants" (20:2).

In scoring, there is no such thing as a wrong answer since each person's response is peculiar to him and reflects his own way of thinking and feeling. Responses have to be scored and analyzed in an individual manner, in such a way that a trustworthy assessment of personality might take place with a reasonable degree of reliability.

At present, little can be said for the reliability of the TAT, as responses reflect the mood as well as the present life situation of the subject. Murray indicates that the reason for this might be that:

> TAT stories offer boundless opportunities for the projection of one's own complexes or pet theories, and the amateur psychanalyst who is disrespectful of solid facts is only too apt to make a fool of himself if in interpreting the TAT, he gives free run to his imagination. The future of the TAT hangs on the possibility of perfecting the interpreter (psychology's forgotten instrument) more than it does of perfecting the material (20:6).

As to reliability, Harrison and Rotter, using a three-point scale (+=1, ?=2, -=3) found a correlation of .73 between two raters, and using a five-point scale (++=1, +=2, ?=3, -=4, --=5) found a correlation of .77 between two raters. On both studies, five pictures were projected on a screen for thirty seconds, and for seven and one-half minutes subjects wrote a story about what they had seen (13:97-98).

III. METHODS OF SCORING TENSION INDICES

The following deal exclusively with the reliability of the DRQ method of measuring tension in different settings.

Dollard and Mowrer's DRQ. The Committee on the Institute of Welfare Research of the Community Service Society of New York wished to have a study made of the nature, cost, and results of the casework process as applied at different times with varying types of clients. Dollard and Mowrer took up one part of the total problem, that of measuring tension changes in a social casework record. These two researchers thought it possible that the pattern of tension movement might be related to "Progress" or "Success" of a case. If high tension levels were seen at the end of the case, it might suggest failure or the necessity of reopening the case. If tension had fallen rapidly, this might be correlated with what the client learned or the value of the Community Service to him. With these two possibilities in mind, Dollard and Mowrer felt it worthwhile to attempt to work out a tension measure. The formula adopted was patterned after the one Binet found successful for measuring intelligence: Mental Age over Chronological Age = I.Q.. The total number of "discomfort" words on a page of a case study was counted. Only words which would stand alone, out of content, as indicating drive-tension were included. The total number of "reward" or "relief" words was similarly determined. From this a quotient was derived by taking the client's total "discomfort" words divided by the total "discomfort" and "relief" words combined.

$$\frac{\text{Discomfort Words}}{\text{Discomfort plus Relief Words}} = \text{D.R.Q.}$$

In practice the index should go up when the reader of the case feels that things are going badly and should go down when matters are going better.

The next major item in their study was to check for reliability of their scorers. The Cellini case, 37 pages arranged in random order, was put to the test. Eight different readers scored this case and the average intercorrelation for the eight curves was +.80. With the success of the word-scoring method as a stimulus, the same method was applied to the sentence. Drive-arousing sentences were scored minus; drive-reducing sentences were scored plus; the neutral sentences were scored zero. The DRQ was computed as described. The average intercorrelation for these eight curves was +.81.

A number of the scorers objected to the sentence scoring method because quite frequently several propositions are compacted into a single sentence in a casework record. Because of this objection the researchers felt that an "independent clause" or "complete thought" might be used as a method of scoring. Ten new scorers were trained thoroughly in scoring by this method. The result of the average intercorrelation of the ten curves was +.88. From these results, the thought-unit method of scoring is superior in that it eliminates certain sources of confusion inherent in the other two methods.

The correlation between the 37 pages of the Cellini case study as to word curve and sentence curve was +.90, between word curve and thought curve +.93.

The problem arose as to which type of scoring yields the greatest degree of coincidence between scorers, another way of looking at reliability. A formula was obtained for determining the degree to which a group of curves actually tends to coincide. When this method is applied to word scoring, a value of 0.01000 is obtained; to sentence scoring, 0.00786; and for clause scoring, 0.00569. The greatest degree of coincidence is in the smaller value.

Dollard and Mowrer found that the average intercorrelation for the eighteen judges was +.64. These results raised a question: How can the DRQ scores be so relatively consistent (with a cofficient of reliability of +.88) for different scorers when the identification of the items which are scored has a reliability coefficient of only +.64? They explained this by stating that the DRQ is a ratio obtained by dividing the total number of pluses and minuses on a given page into the number of minuses. This means that the absolute number of pluses and minuses recorded on a given page does not affect the DRQ for that page. The DRQ of .50 could be the result of 75 pluses and 75 minuses or 25 pluses and 25 minuses.

The original researchers feel that the DRQ has many limitations. They feel that it just measures drive, primary or secondary, single or summated, continuous or serial, that of a client or of any other individual; it is a record of all the tension that creeps into the case record.

They go further to add that the tension index itself does not give clear evidence that learning has taken place in a client. It does, however, picture whether or not there was a chance for learning to take place.

The authors do not claim that the tension index will measure success, but on the other hand they cannot be certain that it does not. It remains a problem for empirical investigation (6:3-32).

DRQ vs Positive Negative Ambivalent Quotient (PNAvQ). In 1948, V.C. Raimy reported a study condensed from a more detailed dissertation which purported to show that by measuring changes in self-evaluation in counseling interviews, the progress of counseling could be traced quantitatively. Although derived from different theoretical positions and applied to different kinds of material, both methods of analysis have the same goal of measuring changes in personality (22:153-159).

Raimy's study is a report of an investigation which attempts to study the relationship between the results of the two methods when both are applied to the same verbatim protocols of counseling interviews.

PNAvQ involves: "P"--Positive Self-Reference; "N"--Negative Self-Reference; "Av"--Ambivalent Self-Reference; "A"--Ambiguous Self-Reference; "O"--No Self-Reference; and "Q"--Non-Rhetorical Questions.

The PNAvQ is then obtained, like the DRQ, by dividing the number of "N" units plus the number of "Av" units by the number of "N" units plus "Av" units plus "P" units:

$$\frac{\text{"N" units} + \text{"Av" units}}{\text{"N" units} + \text{"Av" units} + \text{"P" units}} = \text{PNAvQ}$$

The range of possible quotients, like that of the DRQ, is between and including 0.00 to 1.00 with a quotient closer to 0.00 indicating greater self-approval or, in DRQ terms, greater relief from "tension."

Using analysis of variance, twelve of the thirteen interviews analyzed individually recieved F values which were significant beyond the one per cent level of confidence while one F value was significant beyond the five per cent level. These results permit rejection of the hypothesis that positive and negative self-references represent the same population of DRQ values and thus associate a low DRQ value with the positive self-references and a high DRQ value with the negative self-references in those commonly scored units.

The raw score form of the rank difference method of correlation using the two quotients obtained for each of the seventeen interviews resulted in a Rho of +.838.

Despite the differences revealed by a comparison of the two methods of studying changes in personality during longitudinal contacts, it would appear that both methods produce essentially similar results. The DRQ can be used for almost all types of documents: autobiographies, TAT protocols, or other non-conversational documents. The PNAvQ depends upon the interruptions provided by two people in conversation with each other.

The notion of "Tension" is primarily a nonphenomenal variable used as an explanatory construct for dealing with a wide variety of observed behavior. But in their instructions for discovering the "Tension Condition" of a subject at any given moment, Dollard and Mowrer depend heavily upon "Discomfort-Relief," both terms which refer primarily to the phenomenal realm. Thus, although the tension factor is basic to their theory, Dollard and Mowrer's operations would probably produce almost identical results with Raimy if the underlying hypothesis were ignored.

Raimy's study therefore throws little light upon the validity of the "Tension" hypothesis although it does indicate that results of the DRQ procedure can be interpreted by means of other constructs which do not refer to tension.

Both DRQ and PNAvQ seem to trace in a similar fashion changes from maladjustment to adjustment. Both make use of operations which deal with subjectively reported experiences (22:153-163).

DRQ vs Clinical Ratings. In a study financed by a research grant from the U.S. Public Health Service, Arnold

Meadow reported that the DRQ was being used to such a degree that the question was raised whether it is measuring what it purported to measure. The basic assumption of this method is that the conscious verbal expression of discomfort reflects more or less accurately the psychological tension characterizing the subject (18:658).

Clinical experience suggests that this assumption may be invalid. In some cases the content of the patient's verbalizations appears to reflect the tension noted in behavioral and physiological observations; in other cases there appears to be no relationship between behavioral and physiological estimates of tension and the patient's verbalization.

Meadow's studies were designed to test the validity of the DRQ as a measure of tension and adjustment.

The first study concerns the relationship between the DRQ and the clinical ratings of tension. The method consisted in deriving a DRQ for each of thirty-five chronic schizophrenic patients from a free verbalization protocol and correlating the scores so obtained with a rating of tension made by a psychiatrist.

In obtaining the DRQ, each unit is classified as showing (1) discomfort or dissatisfaction, (2) comfort or satisfaction, and (3) neutrality.

The test was administered to each patient twice within a two-week period to determine reliability. Rank order correlation of test-retest reliability was +.82. The rank order correlation between two independent scorers of the same protocols was +.77. The psychiatrist's rating of tension compared to the clinical psychologist's ratings showed a correlation coefficient (between the two) as +.78.

The correlation between the psychiatrist's clinical rating of tension and the DRQ derived from the free verbalization protocol was +.05. This data indicate no direct relationship between the verbalization of tension and judged overall clinical tension level of the patient.

The second study consisted of correlating the DRQ with a battery of tests designed to measure the degree of personality adjustment of the same thirty-five schizophrenic patients. The DRQ for each patient was the same one described in the first experiment.

Three types of measures used to appraise "adjustment" are paraphrased in outline form as follows:

1. Test of Abstract Thinking
    a. The Neutral Proverbs Test
    b. The Emotional Proverbs Test
    c. The Neutral Similarities Test
    d. The Emotional Similarities Test
    e. The Object Sorting Test

2. Measure of Looseness of Association
(Free verbalization protocols)

3. Rating of Personality Integrations
(Combined clinical criteria of social withdrawal, divorce of thought content from affect, and deterioration of habit pattern).

Test-retest reliability coefficients for the measures of Abstractions ranged from +.89 to +.94; for the Looseness of Association Index the test-retest reliability coefficient was +.78. Interscore reliability coefficient for the test of Abstraction and Looseness of Association ranged from +.89 to +.97. All correlations were significant at the one per cent level of confidence.

Table I taken from Meadows study shows correlations between DRQ and measures of adjustment and indicates positive correlation between all of the measures of abstraction and DRQ. These correlations are all significant at the five per cent level of confidence with the exception of the correlation between DRQ and Object Sorting Test, where p = .06 (18:658-660).

A positive correlation exists between DRQ and Rating of Personality Integration and a negative correlation with Looseness of Association. This shows an indication that the greater the amount of discomfort expressed, the less is the impairment in abstractions, looseness of association, and personality disintegration in this type patient. The results

TABLE I

CORRELATION BETWEEN DRQ AND MEASURES OF "ADJUSTMENT" OF SCHIZOPHRENIC PATIENTS AS REPORTED BY ARNOLD MEADOW

Measures	Number of Cases	Rank Order Correlation	Probability
DRQ and Natural Proverbs	28	.57	.01
DRQ and Emotional Proverbs	27	.41	.05
DRQ and Neutral Similarities	31	.48	.01
DRQ and Emotional Similarities	31	.61	.01
DRQ and Objecting Sorting	20	.31	.06
DRQ and Rating of Personality Integration	31	.47	.01
DRQ and Looseness of Association Index	30	-.48	.01

indicate a positive relationship between the DRQ and all measures of adjustment.

The results of both experiments are interpreted to indicate that a relatively high DRQ cannot be used as a valid measure of tension but may be used as an indicator of good adjustment in schizophrenia.

The assumption of previous investigators that relatively low DRQ is indicative of low "tension" state, good adjustment, and therapeutic success is challenged by the results of this experiment (18:660-661).

DRQ--Dictated vs Verbatim Interviews. At the 1950 American Psychological Association meeting, L. S. Kogan presented the findings of his investigation of the degree of correspondence between DRQ's derived from a dictated interview records and the DRQ's derived from the verbatim interviewee statements in the same set of casework interviews. Reliability between judges was reported as varying from +.80 to +.88 (17:237).

The product-moment r was then computed between the corresponding pairs of DRQ's for the thirty-six dictated and verbatim interviews. This r of +.64, although significantly different from r of zero, is not remarkably high. This degree of correlation might, nevertheless, warrant the use of dictated records for obtaining DRQ in a large-scale study of DRQ for casework interviews.

The question of the utility of the DRQ as a measure of the effectiveness of therapy has been explored in Kogan's study. Although apparently highly rated to self concept ratings in non-directive counseling interviews, DRQ difference scores based on analysis of the initial and closing periods for given cases were found to have a rather low correlation with caseworker judgments of "movement" in the same cases. Only further study will reveal the usefulness of this type of approach (17:236-258).

DRQ--progress with no drop in tension. In a study conducted through a grant from the U.S. Public Health Service, E. J. Murry, supported by Dr. John Dollard, found that a case may show progress even though there is no drop in verbal tension. The purpose of the research was to show that a psychotherapy case in which there was no drop in tension might show progress in other ways. The DRQ was used as a measure of verbal tension, and a content analysis in terms of motivation and conflict was used to show other verbal changes (19:349).

Two separate applications of the DRQ in this case showed no change in verbal tension during therapy. Each application was demonstrated as being reliable (r of +.92 and +.96).

It was pointed out that the DRQ may not reflect the tension experienced by a patient. It was also pointed out that although an eventual tension reduction in a successful patient's everyday life would be expected, a reduction of tension in the therapeutic situation would not necessarily be expected.

The author concludes that although the DRQ may be a good measure of the tension expressed in patient's sentences, a study of this case provides no support for the view that the DRQ is useful for assessing therapeutic progress. In this particular case, other measures of the content of the patient's speech proved to be more promising than the DRQ (19:349-352).

DRQ vs validity. A question has been raised by Frank Auld and associates as to the validity of the DRQ as a measure of tension. The author compared DRQ scores of thirty-nine psychiatric interviews with the global ratings of the same interviews on anxiety, hostility, and dependence. Inter-judge reliability ran from +.22 to +.68. There were small positive correlation between the global ratings and the DRQ, these correlations tending to be higher for the twnty-six women than the thirteen men. If the global ratings can be accepted as adequate measures of tension, the DRQ does not measure individual differences in tension.

However, one must keep in mind that this study focuses on measurement of individual differences in tension and that it is not immediately pertinent to use the DRQ as a measure of change in tension within a series of interviews with an individual patient (2:386-388).

Summary of DRQ literature. One of the most striking observations in this review of the DRQ literature was the great number and degree of contradictory results obtained by different investigators. It seems evident, then, that additional research is necessary as to (1) the reliability of the DRQ as a method in measuring tension and (2) its potential as a possible predictor of human performance. This might serve as one of many research dimensions in order that better and more efficient prediction might be made possible for guidance and counseling purposes.

CHAPTER III

METHOD OF APPROACH

As stated earlier, the purpose of this research was to investigate the scorer reliability of the Thematic Apperception Test (TAT) scored by the DRQ method of measuring tension, a subordinate aim being to investigate some of its interrelationships with certain variables in the prediction of educational success.

I. TESTING PROCEDURE

In the Autumn Quarter, 1954, a group of 134 freshman and sophomore students enrolled in four of the five General Psychology Classes at Central Washington College of Education (CWCE) were asked to take part in this investigation. The students of these four Psychology Classes were given an explanation of the general nature of the study and asked to place their names and place of residence at the top of each page. The students were asked to place their names on each page for follow-up purposes, and this information also was used in another research problem.

Six TAT cards were individually projected by the use of an opaque projector on a 60 x 60 inch screen in a darkened classroom for thirty seconds each. Following each picture, students wrote stories elicited by the pictures. The six TAT cards projected were:

Picture #1. A young boy is contemplating a violin which rests on a table in front of him.

Picture #2. Country Scene: in the foreground is a young woman with books in her hand; in the background a man is working in the fields and an older woman is looking on.

Picture #4. A woman is clutching the shoulders of a man whose face and body are averted as if he were trying to pull away from her.

Picture #10. A young woman's head against a man's shoulder.

Picture #13 M.F. A young man is standing with a downcast head buried in his arms. Behind him is the figure of a woman lying in bed.

Picture #18 G.F. A woman has her hands squeezed around the throat of another woman whom she appears to be pushing backwards across the banister of the stairway (20:19-20).

There were no real criteria for selecting these particular cards except that (1) Murray had regarded them as applicable to both sexes and (2) the investigator felt that these pictures would produce the needed protocols for the investiagtion since they are generally more structured and definitive than cards selected more heavily from the latter part of the series.

Following the thirty-second viewing, the lights in the classroom were turned on and the subjects were given six minutes to write a story about the picture just projected. The following directions were read and also written on the blackboard:

> Tell what has led up to the event shown in the picture, describe what is happening at the moment, what the characters are feeling and thinking, and then give the outcome (20:3).

For students who were absent during class administration of the cards, a similar procedure was used individually using the cards themselves without the use of the opaque projector. This meant that the 134 subjects each had six separate pages, one for each of the TAT cards.

II. SCORING

All protocols were then checked for length and scorability by the investigator. Of the 134 subjects, seventy-three were lengthy enough to be scored, at least six sentences. If one story out of the six, for any one individual, happened to be unscorable because of briefness, the complete series for that person had to be eliminated. The protocols for the 73 subjects were then coded. Each class was designated with a capital letter, A,B,C, etc. Each student within that class was assigned a number, and each of the six pages for each individual was assigned a small case letter so that the code might read A-1-a for the first picture viewed for the number one person in class A. Class A contained twenty subjects; class B contained twenty-one subjects; class C had thirteen subjects; class D contained fifteen subjects; and class E, the students absent for the class administration and given the individual procedure, contained four subjects.

In the Dollard and Mowrer investigation, there are three different methods of scoring: Word, Clause or Thought Unit, and Sentence. They reported a reliability of +.81 using the sentence scoring method (6:12). Reliabilities of +.30 to +.96 have been reported by other investigators using sentence scoring as well as other mentioned methods (25:4).

Since reliabilities can be moderately high for the more rapid sentence method and because of the heavy class load of the graduate students who served as scorers, the sentence scoring method was used in this investigation. The writer numbered all the sentences in the 438 protocols so that each scorer would have the same number of responses.

Instructions on how to score the stories were given to each of the scorers. Practice sessions were conducted, and the stories rejected for brevity were used in the practice runs. Each scorer was told to read each sentence in terms of whether the information it contained indicated (a) Discomfort, (b) Relief, or (c) Neither. The scorer reads each sentence in terms of whether the information it contains (a) disturbs, (b) relieves, or (c) fails to affect him decisively one way or another. The scorer should rehearse the sentence to himself; as he does so, the sentence produces tension or relaxing responses in the scorer, or he experiences no change in tension level. Further details as to the scoring techniques can be found in Appendix A.

When it was felt that the method was clearly understood by the scorers, the 438 protocols were distributed in thirds, one-third going to each of the two scorers and one-third to the investigator. The protocols were rearranged so that they were in random order. When the stories had been scored independently by each of the scorers, they were interchanged until all protocols were scored by all scorers.

For scoring the stories each scorer was given five master scoring sheets, (Appendix B) one for each of the five sections. These master sheets were divided into six sections, one section for each of the six protocols supplied by the 73 subjects. Each of these six sections was subdivided into thirds with headings of Tension, No Tension, and Neutral. If the sentence inferred Tension, then a small mark was placed in the Tension division, if it suggested No Tension, a mark was placed in the No Tension section. This was continued until the story had been completely scored. Then the marks were counted in each division and a large number placed in that subdivision. The scorer then went to the next story. The protocols were strictly scored from code.

When the scorers had completed all the stories and the totals for Tension, No Tension, and Neutral were tabulated, the master sheets were returned to the investigator.

The DRQ's were not computed and compiled at this time. Fortunately, for sake of later follow-up, the investigator entered professional employment and was unable to complete the reliability study at that time.

III. FOLLOW-UP

The study was resumed in the Autumn Quarter, 1960. One reason for the delay was that the investigator had to allow at least five years so that those students who were freshman could complete the four years of college and possess one year of teaching experience. This was important because one problem of this research was to investigate the first year teaching evaluation made on these new teachers by their school district supervisors, in order to study their teaching success.

The next step in the investigation was to obtain from the Office of the Registrar at Central Washington College of Education the transcripts for the 73 students for whom scoring of TAT protocols was accomplished. The information taken from these transcripts was (1) Cumulative Grade Point Averages, (2) Number of Quarters Completed at CWCE, and (3) Grades Received from their Student Teaching Assignment. All transcripts with the exception of one were located and the necessary information procured.

Of the seventy-two subjects whose transcripts were located, twenty-seven completed their student teaching assignment and received Bachelor of Arts Degree in Education. Three additional students received Bachelor of Arts Degree in Arts and Science, two in Economics, and one in Mathematics.

The First-Year Teaching Evaluations were on file in the Office of the Dean of Graduate Study. These records showed that only fourteen first year teachers of the twenty-seven who had received degrees had been evaluated on rating scales. This seems like a very small number; however, of the twenty-seven that graduated, twenty were women, and some married and did not go into the teaching field. A second reason is that some of the teachers left the State of Washington, and the law requires only evaluations of those teachers that remain in the state. Thirdly, some of the male graduates with service commitments may not have entered the teaching field as yet.

A further investigation showed that none of the original seventy-three subjects had been dismissed from school because of disciplinary reasons or because of academic failure.

IV. STATISTICAL APPROACH

After all the data had been gathered, the first step was to compute the seventy-three subjects' DRQ scores for each of the three scorers. The results of the reliability of the scorers were computed using Pearson Product Moment Correlation Coefficient, as were the interrelationships between DRQ and Grade Point Averages, Number of Quarters attended at CWCE, Practice Teaching Ratings or Grades, and First-Year Teaching Evaluations. These will be found in the next chapter.

CHAPTER IV

RESULTS

Investigating scorer reliability of the DRQ method of scoring TAT protocols was the primary purpose of this investigation. A secondary purpose was to investigate some of its interrelationship with certain variables in a teacher education program.

After all the data were tabulated, Pearson Product Moment Correlation Coefficients (r) were computed to discover the reliabilities that existed between scorers. Correlations were also calculated to discover if a relationship existed between the DRQ method of scoring TAT protocols and college success as well as teaching success. The formula for these correlations as well as an example can be found in Appendix C.

I. RELIABILITY BETWEEN SCORERS

Each scorer, graduate students in counseling and guidance, was coded with a letter: X, Y, and Z. Correlation coefficients were calculated between each of the scorers to determine reliability coefficients that might best be considered coefficients of scorer consistency. Results are shown in Table II.

TABLE II

CORRELATION COEFFICIENTS SHOWING EXTENT OF SCORER CONSISTENCY USING DOLLARD AND MOWRER'S DRQ APPLIED TO SIX TAT STORIES

Scorer	Y	Z
X	.44**	.68**
Y		.27*
Average r (using Fisher's z)	.48**	

** Significant at the One Per Cent Level of Confidence.

* Significant at the Five Per Cent Level of Confidence.

The relationship between scorer X and scorer Z showed a moderately high reliability with an r of +.68. The r of +.44 indicates moderate agreement between scorer X and scorer Y. Relationship between scorer Y and scorer Z showed a moderately low correlation with an r of +.27. However, this was significant at the five per cent level of confidence.

The average r of +.48 (using Fisher's z) for the three scorers indicated a moderate relationship significant at the one per cent level of confidence.

These reliabilities are reasonably consistent with earlier studies of the DRQ scoring method on social casework records. This, of course, is the first known attempt to apply this method of measuring tension in TAT stories. The findings show, at least, that the method of scoring can be done reliably.

Since the reliability of scorer X with Z was found to be highest, it was judged that a combination of their scores would tend to at least qualitatively add to the reliability as a predictive factor in other aspects of this study.

II. VALIDITY AND INTERRELATIONSHIPS

DRQ and college success. The average DRQ obtained from scores assigned to the six TAT stories by readers X and Z formed the predictor variable regarded, by logic and some empirical evidence, as a Tension Index. This index formed the basis for determining the relationship that exists between tension and college success as well as teaching success. These in turn were measured by (1) Grade Point Averages, (2) Number of Quarters Spent at CWCE, (3) Practice Teaching Ratings, and (4) First-Year Teaching Ratings.

All seventy-two subjects were included in this part of the investigation, since DRQ and Grade Point Averages (GPA) were now available for each of the subjects. Using the Pearson Product Moment Coefficient of Correlation once again, the r between DRQ and GPA was found to be —.265. This low minus relationship was found to be statistically significant at the five per cent level of confidence.

The minus, of course, suggests that within the limits of relationship there is a tendency for greater tension to be associated with lower GPA and lower to be more characteristic of the higher GPA students.

While working with these data, a sub-hypothesis was formulated. Perhaps the relationship is not best expressed as linear of straight line function. Perhaps a more complex curvilinear function is involved. A scattergraph approach to the calculation of r gave an inspectional indication that an even more prominent tendency for moderate tension (average DRQ's) was associated with high GPA (Appendix D).

At this point it seemed advisable to apply the computation of a correlation for the regression of GPA, which we shall call (Y), and on DRQ, which we shall call (X), to determine whether this curved regression actually prevailed. The correlation ratio for regression of GPA on DRQ resulted in an eta coefficient (η_{yx}) of .578. Appendix E illustrates the calculation of this problem. The standard error of the eta coefficient was found to be ±.08. This standard error, of course, suggests statistical significance well beyond the one per cent confidence level. These findings suggest that a moderate amount of tension indicates best grades received by the subjects in college and measures of lower amounts of tension as well as greater tension is associated with lower GPA's.

Another possible indication of success in college was thought to be number of quarters completed at CWCE. It is granted that this criterion would be hard to defend as a measure of success, but its prediction could be useful if found possible. The correlation between DRQ and the number of quarters spent at CWCE resulted in an r of +.0037. This low correlation shows that these two measures are unrelated in this study. There is no relationship between tension as indicated by DRQ and number of quarters of college completed at CWCE.

DRQ and teaching success. In dealing with the correlation between DRQ and student teaching ratings or grades, the grades had to be assigned a numerical value in order to compute the r value. In most American colleges the value for the letter grade A is 4.00, B is 3.00, C is 2.00, etc. With this substitution of numbers for letter grades, and with only twenty-seven of the original seventy-two students having completed their student teaching assignment, the corresponding DRQ's for these twenty-seven students along with their student teaching grades were correlated and found to have an r of —.226. With only twenty-seven subjects the r value was not found to be statistically significant. The Tension measure is unrelated to student teaching evaluations in this small sample.

In computing the correlation between DRQ and First-Year Teaching Evaluation, it was found that of the twenty-seven students who had completed their student teaching assignment and had graduated with a B.A. Degree in Education, only fourteen had received first-year teaching evaluations. The corresponding DRQ's for these fourteen subjects were correlated with the evaluations which contained a summary rating scale ranging from Superior (five points) to Unsatisfactory (zero points). An r of +.818 was found for this part of the investigation. Even though only fourteen subjects were involved, it is statistically significant at the one per cent level of confidence. To the extent that one is able to extrapolate practical significance from statistical significance, on fourteen cases there is the suggestion that the higher the tension, as measured by DRQ, the higher the principal or supervisor tended to rate first year teaching success.

CHAPTER V

DISCUSSION

Research in comparitively new or unique dimensions of assessing human personality and predicting future performance leads to possible ramifications or speculations that can best be labeled "Discussion," since ideas supercede results of previous and current research. In this study of reliability and predictive validity of the DRQ method of measuring Tension shown in TAT stories, some comparison of results with the few studies made earlier form a basis for initial aspects of the discussion.

I. MEASURES OF RELIABILITY

This study found that reliability between scorers ranged from +.27 to +.68, with an average r of +.48 (using Fisher's z). In the Dollard and Mowrer investigation a coefficient of reliability of .81 was found using the sentence scoring method, the same method used in this study (6:12). Kogan's findings to the problem of investigating the degree of correspondence between DRQ derived from a set of dictated interview records and DRQ derived from verbatim interview statements in the same set of casework interviews with an r varying from +.80 to +.88 (17:237). Murry's study reported an r of +.92 and +.96 when investigation change in tension under therapy (19:350). The Auld study which compared DRQ scores of thirty-nine psychiatric interviews with interviews on anxiety, hostility, and dependance, found that reliability between judges ran from +.22 to +.68 (2:387).

From the above it is obvious that reliability differs in each study. It is true that each investigator is attempting to solve a different problem, but most desirably, reliability between socrers or judges should remain relatively high in order to secure best results in validation studies.

In this present study, it would have been much more desirable to have had more judges. In using only three, reliability could be either very high or very low. More judges would add greatly in determining what is the most likely reliability, should this predictor variable be used further.

Although reliabilities in this study are comparable to some previous studies of reliability of DRQ, it should be mentioned explicitly that there are no other studies with which to make a direct comparison. Earlier studies have applied DRQ to different forms of verbal protocol. This is the first known study of DRQ applied to TAT stories. The only other study of DRQ applied to projective test data is the DuFresne study using CWCE subjects and advanced student scorers. Results are consistent between these two studies in that most all reliabilities are statistically significant and moderately high, suggesting that the DRQ method of scoring projective data for "Tension" can be reliable.

Another point that should be made in this segment of the study has to do with scorer instructions. It was felt by the investigator and also the scorers that more time should have been given to practice sessions so that scorers would be positive as to method and procedures they are to follow. In this study, instructions on scoring were given and it was felt that the judges were basically prepared, but after scoring was completed, it was felt that more time and practice should have been spent in DRQ scoring procedures. This could be one reason why reliability is not consistently higher. If scorers are unsure of the procedure, it will naturally result in lower agreement or reliability.

Perhaps the most significant factor is that even with limited practice in scoring, moderate agreement between graduate student scorers was reached. Normally, any scoring of projective tests requires extended course work in personality theory and projective techniques, with many course prerequisites prior to these. Here a new use for projective test protocols is suggested. Rather than using the TAT only for global personality assessment, it might well be scored on a less sophisticated basis for single or simpler dimensions of personality, i.e., Tension.

The results of this study as well as of the many others dealing with DRQ point out that other research is needed in this area to determine whether the DRQ method of measuring Tension is a reliable one or whether there are other methods that would be more satisfactory.

II. MEASURES OF SUCCESS

The correlation coefficient (r) between grade point average and DRQ resulted in an r of —.265, using Pearson Product Moment Coefficient of Correlation. This minus finding implies that lower tension is associated with higher GPA and higher tension with lower GPA. Because of these results it was felt that the relationship of DRQ to GPA might not be best expressed as a linear function. The scatter graph approach was used as an inspectional means of determining just what type relationship existed. This graph suggested a definite non-linear regression relationship. With this observation as the basis, a correlation for the regression of GPA to DRQ was computed. The correlation ratio showed an eta coefficient (η_{yx}) of .578 with a standard error of ±.08. These results indicate that a moderate amount of tension, average DRQ, indicates highest grades received by students, and that high tension as well as low tension is related to lower GPA.

This eta coefficient of .578 suggests the possible practical significance, since r's between academic aptitude and college grades are usually about the same level. The question arises, then, do these two measuring instruments happen to measure the same thing or do they measure sufficiently different factors to be useful for multiple factor prediction of the criterion? This is a point for further study; if they do measure different factors, which can now by later investigations be found by correlating DRQ and academic aptitude test scores, better prediction of academic success can be made. Further, if counseling services are provided in the college setting, some change in the tension level and, by inference, change in academic accomplishment might well be effected.

This investigation was conducted on the college level. Another question confronts the investigator: Can this method be used successfully in highschool or even in junior highschools as a method of measuring tension to discover those students who, in this respect, are in need of counseling?

Could this method also be used for the educable type child in Mentally Retarded classes for vocational guidance and placement within the community? Is moderate tension essential to effective learning and productivity? Is extreme pressure or lack of tension a deterrent to learning and productive living?

A second interesting computation was that between DRQ and First-Year Teaching Ratings. A correlation coefficient of +.818 was obtained between these two measuring instruments. Even with only fourteen subjects, it was found to be statistically significant. These results imply that the higher the tension as measured by DRQ, the better the administrator's ratings for first year teaching success.

It is interesting to note that when DRQ and GPA were inspected on the scatter graph, it was observed that a moderate amount of tension produced the best college grades. In the correlation between DRQ and First Year Teaching it was found that the greater the tension, the higher the evaluation. A correlation coefficient computed between GPA and First Year Teaching ratings resulted in an r or +.275. This was found to be not statistically significant.

One questions the disparity. Why should moderate tension seem to produce better grades, indicating some cause as well as relationship, yet higher tension be associated with higher teacher ratings? Obviously, size of sample makes one question whether this would hold up, yet the statistical significance suggests that it might. This of course depends on the reliability of the tension measure, stability rather than scorer consistency in this case. This, of course, would be another area for needed study. However, with assumption of reasonable stability it might be speculated that the more tense person may well present the picture of conscientious work, i.e., neatness, promptness, orderliness, conformity to school regulations, etc., and consequently gain higher ratings from school principals than others. Whether or not the student learns more in this setting is another question and points as with so many other studies of "teacher success," towards the need for studies of criteria of "teacher success."

CHAPTER IV

SUMMARY AND CONCLUSIONS

I. SUMMARY

The purpose of this study was to investigate scorer reliability of the Discomfort-Relief Quotient (DRQ) method of scoring applied to the Thematic Apperception Test (TAT). A second aim was to investigate some of the interrelationships between Tension, as measured by DRQ, and (1) Grade Point Averages, (2) Number of Quarters spent at CWCE, (3) Practice Teaching Ratings, and (4) First-Year Teaching success.

In order to determine these relationships, a group of 134 freshmen and sophomore students in four General Psychology classes, in the Autumn Quarter, 1954, were asked to serve as subjects for this investigation. In each of the four classes, six TAT pictures were projected onto a screen for thirty seconds by use of an opaque projector in a darkened classroom. Following the thirty second viewing, the classroom lights were turned on and the subjects were given six minutes to write a story about the picture just projected. By instructions placed on the chalkboard, they were directed to tell what led up to the event, what is happening at present, the characters' thinking and feelings, and the outcome of the story. For students who were absent, the cards themselves were used without the projector.

The protocols for the 134 subjects were checked for length and DRQ scorability by the investigator. It was found that 73 of these protocols were of sufficient length to be scored. After coding was completed to assure anonymity, the investigator numbered each of the sentences in each of the 438 stories to insure that each scorer would have the same number of responses.

Instructions on scoring the stories were given to each reader. Practice sessions were conducted and the scorers were told to read each sentence in terms of whether the information contained within indicated (a) Discomfort, (b) Relief or (c) Neither. When it was felt that the method of scoring was clearly understood, each scorer received one-third of the protocols along with master scoring sheets for each of the groups. These 438 protocols were arranged in random order and scored strictly by code. Each judge scored all the protocols and returned the stories and master scoring sheets to the investigator. The DRQ's were not compiled at this time, as the investigator entered professional employment.

The study was resumed in the Autumn Quarter, 1960. It had been previously decided to use the Dollard and Mowrer DRQ method of measuring tension in written documents. This was the first attempt to apply this scoring method to TAT stories. The investigator next computed the DRQ's for each subject as scored earlier by the three judges. The formula for measuring tension is:

$$\frac{\text{Discomfort Units}}{\text{Discomfort Units plus Relief Units}} = \text{DRQ}$$

Information was obtained from the Office of the Registrar at CWCE as to (1) Grade Point Average, (2) Number of Quarters Completed at CWCE, and (3) Grades Received from Student Teaching Assignment for the 72 subjects with scorable TAT stories. One transcript could not be located. Of these 72 subjects, 27 had completed student teaching assignments and had received B.A. Degrees in Education. First-Year teaching ratings were obtained from the Office of the Director of Graduate Study. These records showed only 14 first year teachers had been evaluated on a rating scale; 27 had received teaching degrees.

After all the data had been collected, the first step was to compute the reliability between scorers. Using the Pearson Product Moment Correlation of Coefficient (r) it was found that reliability between scorers ranged from +.27 to +.68 with an average r of +.48 (using Fisher's z). The original Dollard and Mowrer study showed an r of +.81 using the same scoring method on case work records. The findings show that the DRQ method of scoring tension in TAT stories can be done reliably.

Using the same correlation coefficient formula, the relationship between DRQ and GPA was found to be —.265.

The minus suggests that there is a tendency for greater tension to be linked with lower GPA and low tension with higher GPA. At this point it was hypothesized that the relationship might be more complex than one shown by linear regression, and a scatter graph approach was used as a means of inspection. It indicated that a curvilinear function was more probably involved than a straight line function. A correlation ratio for regression was computed showing an eta coefficient of .578 with a standard error of ±.08, showing definite statistical significance. These findings indicate a moderate amount of tension is associated with better GPA, whereas low or high tension is associated with poorer grades.

When DRQ was correlated with the number of quarters spent at CWCE, an r of +.0037 resulted, which meant no relationship existed between these two factors. In turn, DRQ's relationship to practice teaching ratings was correlated and found to have an r of —.226, which also showed that these two measures were essentially unrelated.

Correlation between DRQ and First Year Teaching Ratings resulted in an r of +.818, statistically significant at the one per cent level of confidence. Because of the small sampling in this portion of the problem, 14 subjects, caution is needed in deciding practical significance. It does suggest, however, that the higher the tension as measured by DRQ, the higher the rating of first year teaching success or that greater tension in a teacher, depending on the stability of the DRQ, apparently combines with performance factors which elicit these higher ratings.

II. IMPLICATIONS

The results of this study might well have important implications for guiding students. The reliability coefficient obtained in this study was not as high as that in Dollard and Mowrer's study, when comparable methods were used. However, the average r of +.48 (using Fisher's z) was found to be significant at the one per cent level. Other studies attempting to solve different problems found reliability between judges to range from +.22 to +.96. An important part of any validity study is to obtain a high degree of reliability between judges in order to secure best results. Since this is the first known study where DRQ has been applied to TAT, it is impossible to directly compare previous studies of DRQ reliability to this measure. The important finding is that the DRQ method of scoring tension in TAT stories can be done reliably.

Another important point that should be mentioned concerns instructions for scorers. Each scorer involved in the investigation must know exactly what procedures are to be followed. A great deal of time and practice should be put into the scoring instructions and procedures. If the study is to produce highest reliability, this is very vital. In the present study it was felt that more time should have been spent in this area. Stability of the measure over a period of time should now be studied.

From the studies of interrelationship between DRQ and measures of college and teaching success, some implications might be drawn:

1. Since the eta coefficient of .578 between DRQ and GPA suggests that moderate tension and high GPA are associated, if through study DRQ is found to be measuring different factors than academic aptitude tests, then the DRQ could well be an important measure to add to the multiple variables for prediction of college success.

2. If moderate tension, as measured by DRQ, holds up as a predictor in college academics, it might well serve as a predictor in public school settings.

3. Special interest is shown in the DRQ as a possible indicator of trainability in the field of Mental Retardation. Tension might well be one of the factors which account for superior learning of some lower ability persons when their measured intellectual superiors are failing to learn or are learning more slowly.

4. This investigation discovered that the relationship between DRQ and GPA indicated that a moderate degree of tension resulted in better grades. When DRQ and First Year

Teaching Ratings were correlated, it showed that the higher the tension the better the rating. The question arises as to why moderate tension should result in best GPA and high tension result in best first year teaching ratings?

The answers to these questions can only be found with additional research in this area. One major need is for research on criteria of success, particularly, teaching success. If moderate tension is shown by other studies to be related to learning, as indicated in this study, this is one further justification for counseling towards modification of tension level as an aid to academic learning and performance. If new studies are undertaken to evaluate present methods of predicting success in colleges and teaching success, then this study has served its purpose.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Anastasi, Anne. Psychological Testing. New York: The Macmillan Company, 1954.

2. Auld, Frank Jr., and George F. Mahl. "A Comparison of the DRQ with Ratings of Emotion," The Journal of Abnormal and Social Psychology, 53:386-88, November, 1956.

3. Barr, A.S., and others, "Measurement and Prediction of Teaching Efficiency," Journal of Experimental Education, 16:203-81, June, 1948.

4. _____ . "Teaching Competencies," Encyclopedia of Educational Research, New York: The Macmillan Company, 1950, pp. 1446-56.

5. Campbell, William G. Form and Style in Thesis Writing. New York: Houghton Mifflin Company, 1954.

6. Dollard, John, and O. Herbert Mowrer, "A Method of Measuring Tension in Written Documents," The Journal of Abnormal and Social Psychology, 42:3-32, January 1947.

7. DuFresne, George W. "An Investigation of the Open-End Sentence Concerning Its Uses in School Guidance Programs." Unpublished Master's thesis, Central Washington College of Education, Ellensburg, 1955.

8. Duncan, Donald H. "An Investigation of the Inter-relationships Among Certain Personality Tests and Ratings and Their Relationship to Academic Success in a Teacher's College." Unpublished Master's thesis, Central Washington College of Education, Ellensburg, 1955.

9. Edwards, Allen L. Statistical Analysis. New York: Rinehart and Company, Inc., 1958.

10. _______. Statistical Methods for the Behavioral Sciences. New York: Rinehart and Company, Inc., 1958.

11. English, Horace B., and Ava Champney English. A Comprehensive Dictionary of Psychological and Psychoanalytical Terms. New York: Longmans, Green and Company, Inc., 1958.

12. Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill Book Company, Inc., 1956.

13. Harrison, Ross and Julian B. Rotter. "A Note on the Reliability of the Thematic Apperception Test," The Journal of Abnormal and Social Psychology, 40:97-99, 1945.

14. Hertz, Wayne S. "The Relationship Between the Teaching Success of First-Year Elementary Teachers and Their Undergraduate Academic Preparation." Unpublished Doctoral Dissertation, New York University, New York, 1959.

15. Hull, Clark L. Principles of Behavior. New York: Appleton Century, 1943.

16. Kauffman, P. E., and V. C. Raimy. "Two Methods of Assessing Therapeutic Progress," The Journal of Abnormal and Social Psychology, 44:379-85, July, 1949.

17. Kogan, L. S. "The Distress Relief Quotient (DRQ) in Dictated and Verbatim Social Casework Interviews," The Journal of Abnormal and Social Psychology, 46: 236-39, April, 1951.

18. Meadow, Arnold. "The DRQ as a Measure of Tension and Adjustment," The Journal of Abnormal and Social Psychology, 47:658-61, July, 1952.

19. Murray, E. J., Frank Auld Jr., and Alice M. White. "A Psychotherapy Case Showing Progress But No Decrease in the Discomfort-Relief Quotient," The Journal of Consulting Psychology, 18:349-53, October, 1954.

20. Murray, Henry A. Thematic Apperception Test Manual. Cambridge: Harvard University Press, 1953.

21. O'Donahue, John. "The Relationship Between Freshman Examination Scores and Academic Success in the Curriculum of Central Washington College of Education." Unpublished Master's thesis, Central Washington College of Education, Ellensburg, 1951.

22. Raimy, V. C. "Self-Reference in Counseling Interviews," The Journal of Consulting Psychology, 12:153-63, December, 1948.

23. Rogers, N. "Measuring Psychological Tension in Non-Directive Counseling," Personal Counselor, 3:237-64, July, 1948.

24. Stein, Morris J. The Thematic Apperception Test. Cambridge: Addison-Wesley Publishing Company, Inc., 1955.

25. Tomkins, Silvan S. The Thematic Apperception Test. New York: Grune and Stratton, 1947.

26. Van Zee, Warren R. "A Study of Professional Course Grades, College Activities, ACE Scores and High School Grades as Related to Success in First-Year Teaching." Unpublished Master's thesis, Central Washington College of Education, Ellensburg, 1958.

27. Young, Barbara L. "Predictive Value of the Washington Pre-College Differential Grade Prediction Tests Used at Central Washington College of Education." Unpublished Master's thesis, Central Washington College of Education, Ellensburg, 1960.

28. Zimmerman, J. "Modification of the Discomfort-Relief Quotient as a Measure of Progress in Counseling." Unpublished Master's thesis, University of Chicago, 1950.

APPENDICES

APPENDIX A

SENTENCE-SCORING INSTRUCTIONS

This set of instructions for scoring sentences in terms of the Discomfort-Relief Quotient closely follows Dollard and Mowrer's original approach.

You will find attached one-third of the TAT protocols, and after you have scored them according to the given directions, pass them on to another scorer and he in turn will pass his protocols to you. Do this until you have completed scoring all of the stories. The stories have all been coded, and attached are the five master scoring sheets. You will find that the master sheets contain scoring sections for all of the 438 protocols. Also you will find that each sentence of the 438 stories has been numbered in order to save valuable time and also to make sure that each scorer has the same total number of responses.

Your task is to read the protocol from beginning to end and score each sentence in terms of whether the information it contains (a) disturbs, (b) relieves, or (c) fails to effect you decisively one way or the other. If you react with tension to a sentence, i.e., if it makes you feel annoyed, excited, appetitive, or apprehensive, put a small mark in the "Tension" section of the master scoring sheet for that particular coded protocol. If you react favorably, i.e., if the sentence gives you a sense of well-being, relaxation, or satisfaction, put a mark in the "No Tension" section of the master scoring sheet for that particular coded protocol. If you react neither favorably nor unfavorably or if you find that your feelings are about evenly balanced, put a mark in the "Neutral" section of the master scoring sheet for that coded protocol. After you have finished each protocol, count the number of marks in each of the sections and place a large number corresponding to the number of marks in that section. Please do this for the "Tension" section, "No Tension" section, and the "Neutral" section for each of the protocols. After you have scored all of the 438 protocols and have counted each of the responses and numbered them for each section, return the protocols and five master scoring sheets to the investigator. You need not carry out any further calculations. The investigator will total the sections and compute the DRQ's for each of the protocols.

The main objective of this research is to determine how closely different persons agree on their scoring of the same material on the basis of the foregoing instructions.

APPENDIX B

MASTER SCORING SHEET FOR TAT PROTOCOLS

Measuring Tension in Written Documents (Score Sheet)

SECTION	Protocol a			Protocol b			Protocol c			Protocol d			Protocol e			Protocol f			Total		
A	T	NT	N	T	NT	N	T	NT	N	T	NT	N	T	NT	N	T	NT	N	T	NT	N
A-1																					
A-2																					
A-3																					
A-4																					
A-5																					
A-6																					
A-7																					
A-8																					
A-9																					
A-10																					
A-11																					
A-12																					
A-13																					
A-14																					
A-15																					
A-16																					
A-17																					
A-18																					
A-19																					
A-20																					

APPENDIX C

DETERMINING RELIABILITY BETWEEN SCORERS

USING

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT (r)

(An Example)

DRQ SCORER X

$N = 73$

$M_x \ .68$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - Mx^2}$$

$$= \sqrt{\frac{35.57}{73} - .462}$$

$$= \sqrt{.487 - .462}$$

$$= \sqrt{.025}$$

$\sigma x = .158$

DRQ SCORER Y

$N = 73$

$M_y \ .65$

$$\sigma = \sqrt{\frac{\Sigma Y^2}{N} - My^2}$$

$$= \sqrt{\frac{32.65}{73} - .422}$$

$$= \sqrt{.447 - .422}$$

$$= \sqrt{.025}$$

$\sigma y = .158$

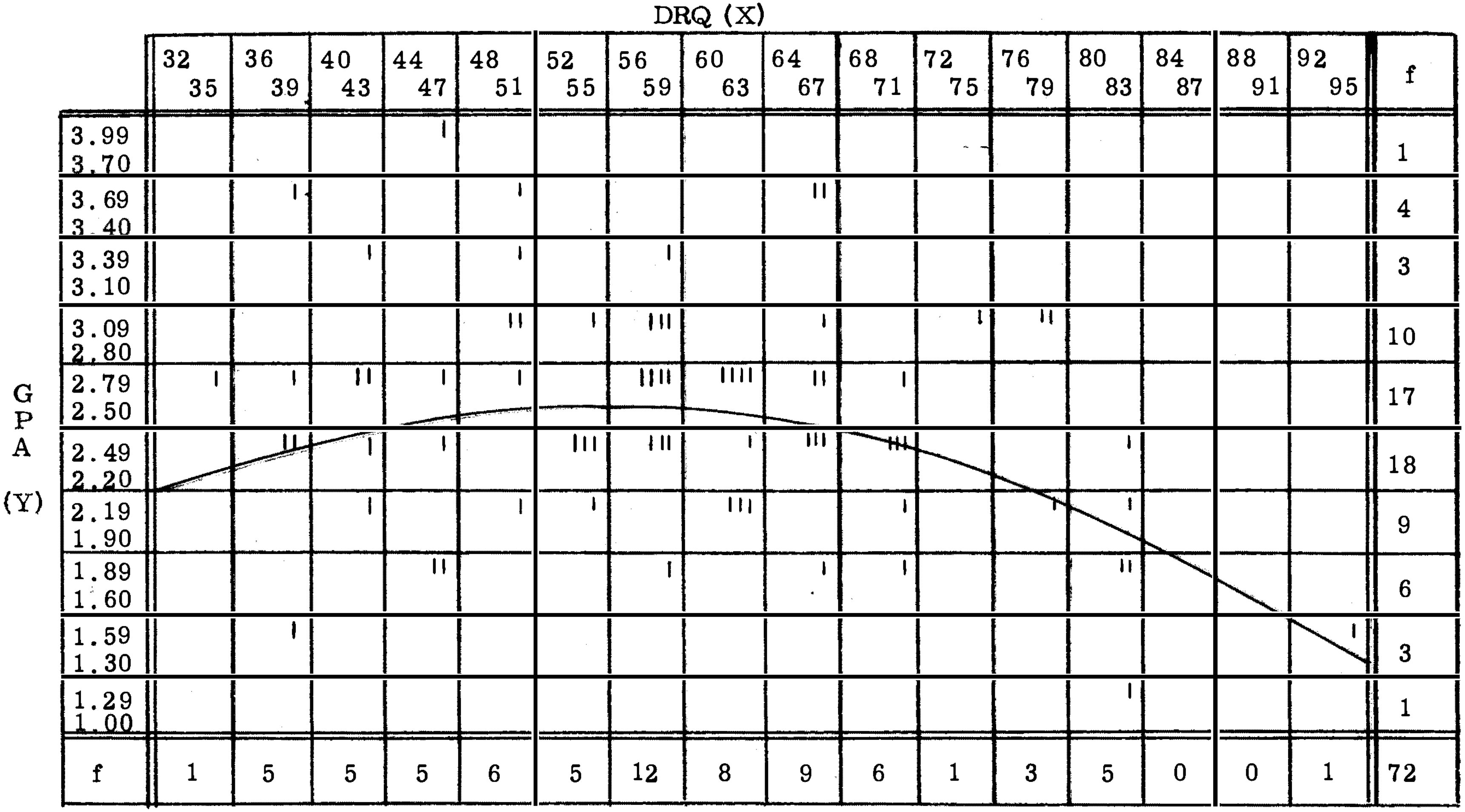
$$r = \frac{\frac{\Sigma XY}{N} - M_x \ M_y}{\sigma x \quad \sigma y}$$

$$= \frac{\frac{33.09}{73} - (.68)(.65)}{(.158) \quad (.158)}$$

$$\approx \frac{.453 - .442}{.0249} = \frac{.0110}{.0249} \qquad r = + .44$$

APPENDIX D

SCATTER GRAPH BETWEEN DRQ AND GPA AND SMOOTH CURVE OF BEST FIT



DRQ (X)

GPA (Y)

	32 35	36 39	40 43	44 47	48 51	52 55	56 59	60 63	64 67	68 71	72 75	76 79	80 83	84 87	88 91	92 95	f
3.99 3.70				I													1
3.69 3.40		I			I				II								4
3.39 3.10			I		I		I										3
3.09 2.80					II	I	III		I		I	II					10
2.79 2.50	I	I	II	I	I		IIII	IIII	II	I							17
2.49 2.20		II	I	I		III	III	I	III	III			I				18
2.19 1.90			I		I	I		III		I		I	I				9
1.89 1.60				II			I		I	I			II				6
1.59 1.30		I														I	3
1.29 1.00													I				1
f	1	5	5	5	6	5	12	8	9	6	1	3	5	0	0	1	72

APPENDIX E

THE COMPUTATION OF A CORRELATION RATIO FOR REGRESSION OF DRQ ON GPA

(1)	(2)	(3)	(4)	(5)	(6)	
X	nc	Y'	Y'-My	$(Y'-My)^2$	$nc(Y'-My)^2$	
93.5	1	1.445	-1.045	1.092	1.092	
89.5	0	0.000	-2.490	6.200	0.000	
85.5	0	0.000	-2.490	6.200	0.000	
81.5	5	1.805	-0.685	0.469	2.345	
77.5	3	2.647	0.157	0.024	0.720	
73.5	1	2.945	0.455	0.207	0.207	
69.5	6	2.196	-0.294	0.086	0.516	
65.5	9	2.345	-0.145	0.021	0.189	
61.5	8	2.382	-0.108	0.011	0.088	
57.5	12	2.620	0.130	0.017	0.204	
53.5	5	2.405	-0.085	0.007	0.035	
49.5	6	2.895	0.405	0.164	0.984	
45.5	5	2.465	-0.025	0.0006	0.003	
41.5	5	2.585	0.095	0.009	0.045	
37.5	5	2.465	-0.025	0.0006	0.003	
33.5	1	2.465	-0.025	0.0006	0.0006	
	72				6.4516	$\Sigma\, nc(Y'-My)^2$
					.0896	$\sigma^2 y'$
					.2993	$\sigma\, y'$

COMPUTATION OF A CORRELATION RATIO--Continued

$$n_{yx} = \frac{y'}{y}$$

$$n_{yx} = \frac{.2993}{.5171}$$

$$n_{yx} = .578$$

STANDARD ERROR OF A CORRELATION RATIO

$$\sigma n = \frac{1 - n^2}{N - 1}$$

$$\sigma n = \frac{1 \ - \ .326}{72}$$

$$\sigma n = \frac{.674}{8.48}$$

$$\sigma n = .08$$

$$\sigma n_{yx} = .578 \pm .08$$