THE MEASUREMENT OF EMOTIONAL HEALTH THROUGH THE USE OF ESTAVAN'S MODIFIED PAIRED COMPARISON TECHNIQUE

Thesis for the Degree of M. A.

MICHIGAN STATE UNIVERSITY

Ross E. Carter

1966

ABSTRACT

THE MEASUREMENT OF EMOTIONAL HEALTH THROUGH THE USE OF ESTAVAN'S MODIFIED PAIRED COMPARISON TECHNIQUE

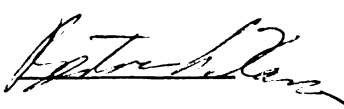
by Ross E. Carter

The purpose of this paper was to determine if judgments of emotional health could be quantitatively measured using Estavan's modified paired comparison method to derive a scale value for each stimulus judged, as well as to assess the reliability of such measurements.

Six protocols of 20 TAT stories each were presented in $\frac{n\ (n-1)}{2}$ pairs to two judges who judged the amount of emotional health of one member of a pair as compared to the other member. The Estavan method of modified paired comparisons was used. This method requires that the member of a pair judged greater on an attribute be represented by a 20 centimeter line and that the lesser member of the pair be compared to the greater by placing a point on the 20 centimeter line which indicates how much, in comparison to the greater member, the lesser member has of the attribute being judged. This procedure results in a ratio or proportional judgment.

Ratio scale values were derived for each set of TAT stories for each judge. A measure of inter-judge

reliability resulted in a correlation of .87. Measures of intra-judge reliability, using a method similar to Gulliksen and Tukey's for Thurstone's paired comparison data, showed that the scale values accounted for .79 of the variance of Judge 1 and .93 of the variance of Judge 2.

Approved: 

Date: _____

THE MEASUREMENT OF EMOTIONAL HEALTH THROUGH THE USE OF ESTAVAN'S MODIFIED PAIRED COMPARISON TECHNIQUE

By

Ross E. Carter

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Department of Psychology

1966

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Dr. Bertram Karon who was most patient and who gave freely of his time.

A special word of thanks goes to the author's wife who did all those good things which often lighten the heavy load of work.

DEDICATION

To Charl

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

INTRODUCTION

The purpose of this paper was to determine if judgments of emotional health could be measured in such a way that ratio scale values on an unidimensional scale could be assigned to stimuli judged, and if the method of judging used was reliable for both inter- and intra-judge comparisons.

As Thorne (1961) points out, judgments of a clinical nature have traditionally been thought of as being based in intuition or other personal, subjective factors. This view has fostered the belief that clinical judgment could not be systematically and rigorously investigated due to its incongruity with objective measures and methods.

Clinical judgment, or rather, judgment in a clinical situation of clinical material, may however, be thought of as not necessarily qualitatively different from other judgments. As Johnson (1955) describes it, judgment is the decisive or end product of an intellectual problem solving activity which has the function of evaluating or settling an uncertain state of affairs. From such a point of view there seems to be little to indicate that clinical judgment differs from any other sort of judgment except in terms of the type and complexity of the material being judged.

As Hunt and Jones (1962, p. 34) says of the comparison of psychophysical and clinical judgments:

> They are merely the opposite poles of a rough continuum, a quantitative continuum marked by the clarity and specificity with which the stimuli are defined, by the degree to which the judgmental setting is standardized through careful control of the known pertinent variables and the elimination of extraneous cues, and by the provision of uniform modes of reporting or response that lend themselves to convenient mathematical treatment.

Research in the area of clinical judgment has not reached the refined point of psychophysical judgment. One difficulty has been the lack of any method for measuring clinical judgment in an exact manner. The importance of this paper seems to lie in the fact that it introduces a method for obtaining a refined measurement of judgments of clinical material which results in a ratio scale value so that differences between the scale values can be interpreted as reflecting actual differences between the stimuli measured.

REVIEW OF THE LITERATURE

A review of the literature on clinical judgment indicates that a major portion of the research in the area has been concerned with showing how the reliability and validity of judgments are effected by certain variables associated with either the materials or the judges.

The results of such research have been contradictory and are confusing. It is suggested that much of this confusion is due to the fact that a refined and accurate method of measuring clinical judgment does not yet exist.

In most cases, clinical judgments have been expressed through either ranking or rating techniques which, in turn, have been analyzed by the use of correlational methods. One exception to this is found in a study done by Albee and Hamlin (1949) where Guilford's (1928, 1931) method of paired comparison preferences was used to obtain scale values for clinical judgments which were then tested for reliability by comparison to ranked orders of adjustment as made by clinicians. While this is a more sophisticated method of measurement than is found in most studies, its use can be criticized on the basis that the number of judges used by Albee and Hamlin was less than Guilford's method requires to produce reliable results.

Even though rating and ranking methods, which in turn, can be studied by correlational methods, suffice in some studies, it should be remembered that rating and ranking methods are subject to errors of leniency, central tendency, and halo effects, as well as anchoring and context effects, and that correlational methods only serve to show associational relationships. Moreover, it is questionable whether the assumptions underlying correlation coefficients are met by such methods let along the assumptions underlying the analysis of variance, "t" tests, and other common powerful statistical tools. It would seem that more exact findings would result if a better method of measuring clinical judgment could be devised and used.

This review of the literature will deal with those studies which have shown clinical judgment to be affected by certain variables such as judges' experience and use of the materials, as well as kinds and amounts of materials. This review will also deal with those studies which have specifically manipulated stimulus properties in order to show an affect on clinical judgment. In order to reduce confusion the research has been divided into sections and will be reported on under those sections.

Experience of Judges

Using schizophrenic responses to vocabulary test items from an intelligence scale as material in a study designed to compare the judgments of experienced clinicians

with those of inexperienced judges, Arnhoff (1954) found that the reliability of the judgments, as expressed on a rating scale, decreased with increases in experience, so that experienced clinicians actually produced less reliable judgments than did the naive or inexperienced judges.

In a follow-up on Arnhoff's study, Hunt, Jones, and Hunt (1957), using a set of improved instructions found that while there was no significant difference between the mean reliability of judgments made by experienced judges and naive undergraduates, there was a significantly smaller degree of variance in the judgments of the experienced clinicians which indicated that reliability, defined as inter-judge agreement, was greater for the experienced judges.

In further investigations on the reliability of experienced and inexperienced judges using rating scales, it has been found that while both experienced and naive judges can make reliable judgments of clinical material (Luft, 1950; Bialick and Hamlin, 1954; Weitman, 1962; and Allison, Korner and Zwanziger,1964), naive judges tend to have difficulty in making judgments which require finer discriminations (Hunt and Jones, 1958a; Hunt and Jones, 1958b).

While it is difficult to account for the ability of naive judges to make as reliable judgments as do experienced clinicians, one explanation could be found in terms of there being various levels of ability among judges in the

experienced groups. Even though Grigg (1958) found no relation between various levels of experience and varying levels of reliability, there is other evidence which does support this notion.

Hunt, Arnhoff, and Cotton (1954) investigating the individual reliability coefficients of experienced judges, using rating scales, found a range of from +.02 to +.93. The results of their study agree with those of a study done by Phelan (1965) who used a matching task to make interjudge comparisons among experienced clinicians for reliability. Phelan found that while all interjudge comparisons in his study were fairly reliable, there was a wide range of reliability. Further evidence for there being varying levels of ability among experienced clinical judges comes from research done by Holsopple and Phelan (1954), Phelan (1960, 1964), and Gunderson (1965).

From these studies it could be concluded that experience in and of itself does not result in increases in ability to make reliable clinical judgments and that in spite of some judges having high ability in the experienced groups, the low ability of some judges operates to equate the reliability of the experienced judges with that of a naive group.

Use of Materials

While experienced judges may vary in terms of the amount of ability in making judgments, another factor which

seems to be involved in clinical judgments is the way the individual clinician uses the materials.

Raines and Rohrer (1955) found that while experienced judges described personality traits of subjects in a way which agreed with external criteria, the judges themselves differed as to what they felt were important traits. The authors concluded that these differences were due to personal factors among the judges which resulted in selective sensitivity to particular elements in the material.

Further evidence for variation among clinicians in the use of materials comes from Golfarb (1959), who found that diagnostic judgments varied with individual clinicians and from Grosz and Grossman (1964) who found significant differences among clinicians in the reporting of anamnestic data which were emotionally charged, as well as from Mehlman (1952) and Pasamanek (1959).

Types of Material

It would seem that, to some extent, judgment should be related to the types of material used in the judgment task. Several studies have investigated the reliability of clinical judgment as it is related to various materials.

While Soskin (1959) found no difference in the reliability of groups of judges using either objective test data, projective test protocols, observations of role playing situations or biographical data, used either alone or in succession, Sines (1959) found that the use of biographical data added accuracy to judgments made only on the basis

of test data.

Kostlan (1954) also studied the effect of the kinds of materials used in clinical judgments by varying kinds of information given to clinical judges and found significant differences as information was varied. In particular, his study showed that predictions from social histories alone were as reliable as predictions from a combination of TAT, MMPI, and Rorschach protocols.

Little and Shneidman (1959) investigated congruences between personality descriptions made by clinicians on the basis of different sources of information such as anamnestic data, MAPS, TAT, Rorschach and MMPI protocols, and found that reliability, defined as agreement between judges, was greater when judgments were based on anamnestic data than when based on any other source of material.

Further evidence that test data alone do not lend themselves to accurate clinical judgments comes from the work of Mancuso (1961) and Horwitz (1962).

Such findings as these seem to indicate that test data do not form an adequate basis from which to make reliable clinical judgments. This is important since tests are often used in actual clinical practice as a basis for judging personality dynamics. We shall have occasion to challenge such conclusions about the adequacy of test data, for, when these results are thought of not in terms of types of material, but rather in terms of amount of material, a new and critical variable seems to be more important.

Amount of Material

Hamlin (1954) reviewed ten studies of clinical judgment which had used projective tests. Five of the studies had shown positive results and five had shown negative results in terms of the reliability and validity of the judgments.

In comparing the amounts of material used, Hamlin hypothesized that when global or atomistic units of information were used, the effect was to produce negative findings. He concluded that it was not the type, but the amount of material used which was important, and suggested that the optimal amount of material to be used was one which was large enough to allow the judges to formulate patterns of the subject's personality, but small enough that the judges were not overwhelmed by the material.

Hunt and Walker (1962) found that valid and reliable judgments could be made from vocabulary and comprehension scales of intelligence tests using a global approach. This would seem to contradict Hamlin's hypothesis except that the reliability of the global approach used in this study may have been due to the limited scope and homogeneity of the stimulus materials.

Jones (1959) investigated the reliability and validity of judgments made from individual intelligence test items as well as from global appraisals of the test protocols, and found that increased amounts of material lowered the

reliability of both the experienced and inexperienced judges, but did not effect validity. In contrast to Jones, Levine (1954) has reported that validity is decreased by increases in amounts of material, but not reliability.

Powers and Hamlin (1957) found that judges who made reliable and valid judgments used several items of information more frequently than they used either one item or all items. Supporting evidence for this finding has been offered by Martin (1958) and Lee and Tucker (1962).

Miller and Bieri (1963) using an information theory approach, studied the channel capacity of clinicians by varying the amounts and the types of information given to judges. Their study showed that about one bit of information was all that could be handled reliably by judges, with some variations due to the type of information and type of judgment involved. One may question the findings of this study on the basis of whether or not these results would generalize to types of material other than those used in the study.

In investigating the use of the total Rorschach protocol, Grant, Ives and Ranzoni (1952) found that reliability of judgments was low when based on the total protocol.

Cummings (1954) found however, that reliability could be achieved when only one Rorschach card was used. Newton (1954), in contrast to both Grant et al., and Cummings, found that reliable judgments could be made using total Rorschach protocols, but concluded that these results were obtained

only because judges were allowed extensive time in which to analyze the protocols.

Thus, either limiting the amount of material or giving the judges sufficient time to assimilate the information contained in large amounts of material resulted in increased reliability. It could be concluded from these studies that the amount of material which is used in research on clinical judgment does affect the reliability of the judgments and that in making judgments of clinical material, there is a limited amount of information which can reliably be handled at a given time.

Stimulus Variables

Only a few studies have attempted to demonstrate the affect of the characteristics of the stimulus materials or methods of presentation on the judgment process itself.

Campbell, Hunt, and Lewis (1957) studied context effects in judgments of adjustment using rating scales, by varying the context in which stimuli were presented and found thatassimilation and contrast effects were produced and caused distortions of the judgments.

Jones (1957) produced context effects in judgments about severity of schizophrenia by presenting a limited range of stimuli, but allowing judgments to made on a full range of pathology.

Context effects have also been shown in the studies of Levy (1960) and King, Ehrman, and Johnson (1952).

Jackson (1963) studied the affects of frequency, extremeness, and order of presentation on clinical judgments and found that extremeness of conflict material was more important than frequency of conflict material in effecting clinical judgments.

Miller and Bieri (1963) in support of Jackson, found that more reliable judgments expressed on rating scales, were made when stimuli were in the extreme ranges of pathology, and that as stimuli decreased in extremeness, so did reliability decrease.

Hunt, Schwartz, and Walker (1965) utilized the results of ratings performed in other studies of Hunt, and found that stimuli rated as extreme in pathology showed smaller deviations and concluded that reliability for these judgments, defined as agreement among judges, was higher than for other stimuli judged less severe.

In Jackson's study mentioned above it was found that judgments of adjustment made from test protocols were affected more by recency of exposure than by primacy. Sines (1959) found that interviews added more to the total reliability of judgments made on the basis of test protocols when judges were exposed to interview material before test data, indicating that primacy effects were greater than recency.

Miller and Campbell (1959) have supplied a clue to the resolution of the conflict over primacy and recency by their finding that neither recency nor primacy effects

are constant during clinical judgment, but depend upon the time at which their measure is taken.

While Arnhoff (1954) was not able to show anchor stimuli caused distortions in clinical judgments, Block (1964) in analyzing his study, found that the judges' personal frames of reference intruded on judgments and exerted a strong anchor effect. Block also noted shifts in frames of reference with changes in the context of the stimuli which is similar to the findings of Soskin (1954).

Block (1962) has also shown that response sets may affect clinical judgments. He devised fictitious test results, and found that deceived clinicians would write clinical descriptions of fictitious patients based on these contrived test results. He concluded that clinical training consists more of indoctrination than of training in the ability to think critically. In contrast to these findings and opinions, Gross (1961) found that response sets had little, if any, affect on clinical judgments. His study showed highly significant stimulus affects in a task requiring the judging of subjects by judges, but little affect due to response sets.

Regarding Block's study, one might legitimately ask what should be expected when clinicians are presented with clinical material and asked to make clinical judgments about the material.

The method of presenting the stimuli to be judged has been shown to have little or no affect on the reliability of clinical judgments. Giedt (1955) and Borke and Fiske

(1957) were unable to demonstrate any affect when clinical material to be judged was presented through direct interview, seeing and hearing interviews, hearing interviews, or reading interviews. Luft (1951) compared the effectiveness of listening with that of reading clinical material and found that judgments in the form of making predictions to responses on objective tests were equal for groups who heard or read the material to be judged, but that prediction of responses to projective tests were more accurately made by listeners than by readers.

In short, clinical judgments are complex and are related to many variables, but the factors affecting them can be investigated.

Many of the findings of research in this area are confusing. It is suggested that this confusion results not so much from the fault of poor research, as it does from the difficulty of dealing with such a complex subject.

It would seem that the complexity of the material demands more rigorous investigation if the subtle factors involved are to be brought to light. One requirement of rigorous research is an exact method of measurement. It is this problem to which this paper is directed.

THE PROBLEM

The purpose of this paper was to investigate a method for measuring judgments of emotional health. The clinical concept of adjustment would seem to be multifaceted yet it is useful to think of adjustment as a single dimension for many purposes. A factor analysis of 14 criteria of adjustment in the Menninger Psychotherapy Project (Luborsky, 1962) showed that 60% of the variance was accounted for by the first principle component, which suggests that much of the variance can be accounted for by a single dimension.

It is also suggested that the data may be more unidimensional than factor analysis suggests because symptom substitution and interchangeability can only be taken into account by a human judge. Therefore, if an appropriate quantitative technique for mapping clinical judgments onto a numerical scale can be developed, it might be found that a single dimension accounts for a surprisingly large amount of the variance. We shall attempt to find out if this is so.

THE METHOD

Sets of 20 TAT stories were obtained from each of six male subjects; two "normal" college students, two college students receiving psychotherapy on an outpatient basis, and two hospitalized schizophrenics. As far as possible, the subjects were equated for age, education, number of siblings in the family and father and mother's occupation. Appendix A lists these variables for the subjects.

Administration of the TAT cards was carried out in standard fashion except that the complete set of 20 cards was administered to a subject at one setting. One examiner was used for all subjects. All subjects were shown the same cards, but not in the same order due to examiner error. Order of presentation is shown in the Appendix.

Stories told by the subjects were first recorded on tape and then transcribed verbatim so that as little distortion or fill-in by the examiner as possible would occur.

The six sets were identified by letter and presented with information regarding the subjects' age, sex and number of siblings to two advanced clinical graduate students

for judgment. Both judges were experienced in the interpretation and evaluation of TAT protocols as well as protocols from other projective devices and, in addition, were functioning as psychotherapists in both group and individual cases.

The attribute to be judged was the emotional health of each subject as compared to every other subject. For the purposes of this study emotional health was defined as being comprised of the following:

(a) Ability to take care of self
(b) Ability to work
(c) Sexual adjustment
(d) Social adjustment
(e) Absence of hallucinations, bizzarre delusions, gross distortions of reality, lack of passivity
(f) Degree of freedom from anxiety and depression, degree of diffuse hostility
(g) Amount of affect, of feelings
(h) Variety and spontaneity of affect
(i) Satisfaction with life and with self, absence of deficiency motivation, i.e., making up for lost love
(j) Achievement of capabilities, mastery of the environment
(k) Benign rather than malignant affect on others

Indications of emotional health as found in TAT stories were defined, in addition, as follows:

(a) Long protocols
(b) Protocols should show more affect, more varied affect
(c) Less stereotyped and more varied material
(d) An increase in benign fantasies and more helping parent figures
(e) Better reality testing
(f) Problems should be directly represented
(g) There should be indications of confidence

Task instructions were given to the judges together as a pair, in both written and verbal form. It was stressed

that while they should use the criteria indicated as guides in forming their judgments, they should rely on their own subjective, clinical impressions and not judge strictly on these signs. The judges were requested to complete a questionnaire regarding the use of the criteria after finishing the task. This questionnaire as well as the written part of the instructions has been included in the Appendix.

At the time of instructions, both judges were given examples of TAT stories representing both extremes of adjustment, in order to show how the criteria of emotional health could be applied to the materials of this study as well as to establish examples of pathology and adjustment as they might appear in TAT protocols. One extreme of pathology was represented by three TAT's taken from hospitalized schizophrenics, while the other extreme was represented by a TAT taken from Wessman and Ricks' (1966) study of college students.

The judges were instructed to judge the TAT stories in pairs, using Estavan's modified method of paired comparisons, so that each protocol was compared to each of the other five protocols. Both judges judged the same pairs independently of each other. For each pair of stories, the judges were asked to judge which member of a pair was healthier, and in comparison to the healthier member, to judge how healthy the other member was.

Each judge was presented with a sheet of paper on

which a 20 centimeter line had been drawn. In expressing his judgment, the judge was instructed that for each pair judged, the healthier member should be represented by the entire length of the line, which should be labeled accordingly. The comparative judgment of the less healthy member of a pair to the more healthy member of the same pair was expressed by placing a point on the 20 cm. line which indicated how much health the less healthy member had, using the emotional health of the healthier member as a standard. This method of comparison results in a ratio judgment.

Comparison of a stimulus with itself, such as (A,A) was not used. Recognizing that reciprocal comparisons such as (B,A) and (A,B) result from the same judgment, there were $\frac{n(n-1)}{2}$, or 15 independent comparisons.

The order of comparison was randomized as is shown in the Appendix, and was carried out so that the protocol listed first in any pair was read before the second protocol.

Since this paper utilized Estavan's modified method of paired comparisons to obtain scale values for the judgments, it may at this point, be useful to describe the rationale for deriving these scale values.

Estavan's Modified Method of Paired Comparisons

For each pair of stimuli judged, the point on the 20 cm. line was measured. The resulting length was divided by 20 to produce a proportion.

The judgments represent the ratio of one stimulus to

another, i.e., the ratio of stimulus B to stimulus A, or B divided by A.

Estavan has found that reliable judgments only occur if the greater stimulus is equated with the fixed length of the line and the lesser stimulus judged as a proportion of the line. When the lesser stimulus was equated with the fixed length of the line and the line extended to indicate the magnitude of the greater stimulus, Estavan found that the judgments were unreliable. Hence, of the judgments, B divided by A and A divided by B, only one can be observed, that one in which the greater stimulus is the denominator. The other judgment can be determined only numerically by taking the reciprocal of the observed fraction.

If we take a hypothetical problem involving three stimuli, A, B, and C, the observations may be arranged in a 3 x 3 matrix (or n x n matrix, where n equals the number of stimuli) as is shown below, with there being a row and a column for each stimulus. The entries in the cells of the matrix are the column stimulus divided by the row stimulus.

	A	B	C
A	$\frac{A}{A}$	$\frac{B}{A}$	$\frac{C}{A}$
B	$\frac{A}{B}$	$\frac{B}{B}$	$\frac{C}{B}$
C	$\frac{A}{C}$	$\frac{B}{C}$	$\frac{C}{C}$

The diagonal entries are by definition equal to one. Half of the off-diagonal entries will be determined by the observations. The other off-diagonal entries are determined by taking the numerical reciprocals as explained above.

Thus, in the comparison of the pair (A, B), A over B will be observed where B is the greater stimulus. B over A is determined from its reciprocal.

It is obvious if we have compared A with B and B with C, that one ought to be able to predict what one would observe if one compared A with C. Such redundancy permits us to observe how well the scaling model fits the data. We shall describe the systematic procedure for doing this below.

To derive the scale values for the observed data, each entry in the matrix of observations is transformed to its logarithm of the base 10 as is shown below.

	A	B	C
A	$\text{Log}\ \frac{A}{A}$	$\text{Log}\ \frac{B}{A}$	$\text{Log}\ \frac{C}{C}$
B	$\text{Log}\ \frac{A}{B}$	$\text{Log}\ \frac{B}{B}$	$\text{Log}\ \frac{C}{B}$
C	$\text{Log}\ \frac{A}{C}$	$\text{Log}\ \frac{B}{C}$	$\text{Log}\ \frac{C}{C}$

The above is equivalent to the following:

	A	B	C
A	Log A -Log A	Log B -Log A	Log C -Log C
B	Log A -Log B	Log B -Log B	Log C -Log B
C	Log A -Log C	Log B -Log C	Log C -Log C

The resulting matrix of differences is at this point, similar to the matrix of differences in Thurstone's Case V Method.

Mosteller (1951) has shown that a least squares solution for the scale values derived from such a matrix of differences is extraordinarily simple. (In our case, it is the sum of squares of errors on the logarithm scale which is being minimized. Although the error term might be defined in some other fashion, this leads to the simplest computational procedure.) One need only sum the columns which yields the following totals:

$3 \log A - \log A - \log B - \log C$

$3 \log B - \log A - \log B - \log C$

$3 \log C - \log A - \log B - \log C.$

If we divide by n, the number of stimuli, we get;

$\log A - \bar{L}$

$\log B - \bar{L}$

$\log C - \bar{L},$

where $\bar{L}$ is the mean of the logs of the scale values. If we set $\bar{L}$ equal to zero (which means that we have chosen the geometric mean of the scale values as our unit of measurement, i.e., 1), then these column averages are our best estimate of the logs of the scale values. Transforming to anti-logs gives us the scale values themselves.

Obviously, any set of judgments, no matter how meaningless could be entered into the matrices and used to derive scale values. One needs some way of evaluating whether the data make any sense, that is, whether the scaling model fits the empirical observations. We are presuming ratings of emotional health can be summarized by a one dimensional ratio scale.

If one has compared stimulus A with stimulus B, and has compared stimulus B with stimulus C, i.e., has a rating of A divided by B and B divided by C, then one can predict what one ought to observe when one compares A with C. If the prediction is correct, then the scale values have summarized the data. If the prediction is inaccurate, the scale values have not summarized the data and the scaling model is inappropriate to the data under consideration.

Gulliksen and Tukey (1958) have presented a procedure for performing such an evaluation over the whole matrix of data. They provide a procedure for dividing the total variance (T) of the empirical observations into two components; variance accounted for by the scale values,

and discrepancy variance (D), variance not accounted for by the scale values. They then define the following index of reliability, R_{ss}, as:

$$\frac{T - D}{T}$$

which summarizes the percent of the variance of the observations accounted for by the scaling procedure. Included in the discrepancy variance are all errors of observation, unreliability of judgment, lack of unidimensionality, and failure of any of the assumptions of the scaling model. Therefore, R_{ss} measures the degree to which the scale values reliably summarizes the data and hence, the degree to which the scaling model is appropriate and valuable.

Inasmuch as Gulliksen and Tukey derived their index for difference rather than ratio observations, i.e., Thurstone's Case V model, R_{ss} can be computed most straightforwardly from the logs of the observations and logs of the scale values.

For Estavan's procedure, R_{ss} measures intra-judge reliability or scalability. To measure inter-judge reliability, one need only compute product moment coefficients as between any other two measuring instruments that yield measurements on a ratio scale.

Analysis of the Data

In the discussion above, a 3 x 3 matrix was used as an example. In the analysis of the data, a 6 x 6 matrix was necessary.

For each judge, a 6 x 6 matrix was determined as above. The $\log_{10}$ of each cell entry was determined to form a matrix of the logs of the observations. The columns of this matrix were summed and divided by 6 to determine the logs of the scale values. By converting these values to their anti-logs, the scale values for each stimulus were arrived at.

In order to determine R_{ss}, a new 6 x 6 matrix of the theoretical observations was determined by subtracting the $\log_{10}$ of the row stimulus from the $\log_{10}$ of the column stimulus. The entries in this matrix represent what the logs of the observed values would have been if the scale values determined were the true scale values and if there were no error variance. Entries in this matrix were subtracted from corresponding entries in the matrix of logarithms of observed values to form the matrix of errors or discrepancies. The entries in half the matrix of errors (either those above the diagonal or equivalently, those below the diagonal) were then squared and added to determine the discrepancy sum of squares. When this is divided by the degrees of freedom of error, $\frac{(n-1)(n-2)}{2}$ the result is the discrepancy variance (D).

The sum of the squares of the entries in half the matrix of the logs of observations (either the entries above the diagonal, or equivalently, those below the diagonal) determines the total sum of squares. When this is divided by the total degrees of freedom, $\frac{n(n-1)}{2}$, or 15, the result is the Total Variance, or (T).

R_{ss}, the intra-judge reliability, as described above, is $\frac{T - D}{T}$. For 6 stimuli, R_{ss} has upper and lower bounds of +1.00 and -.50 respectively.

RESULTS

The scale values obtained for each set of TAT stories are shown in Table 1 for both Judge 1 and Judge 2, and represent the amount of emotional health each subject was judged to have.

Table 1. The scale values for the stimuli judged by two judges.

	Judges	
Stimuli	1	2
A	3.0130	1.8150
B	1.0570	.8823
C	1.6730	1.4580
D	.7171	.4569
E	.2327	.6696
F	1.1250	1.4000

Intra-judge reliability, R_{ss}, for the degree of internal consistency for each judge, was found to be .79 for Judge 1 and .93 for Judge 2. The judges agreed in the designation of the healthier member of a pair in 14 out of a possible 15 cases. The pairs picked are shown in the Appendix.

Inter-judge reliability calculated through a Pearson r correlation coefficient was found to be .87, significant at the .05 level.

Measures of validity were determined only indirectly since this was not a major concern of this paper. The scale values derived for the two "normal" subjects, the two subjects receiving psychotherapy and the two hospitalized schizophrenics are shown in Table 2.

Table 2. Scale values for subjects by class.

	Judges	
Subjects	1	2
Normals		
A	3.0130	1.8150
D	.7171	.4569
Psychotherapy		
C	1.6730	1.4580
F	1.1250	1.4000
Hospitalized		
B	1.0570	.8823
E	.2327	.6696

Inspection of Table 2 shows that, with one exception, the highest scale values began with the normal subjects, decreased through the subjects receiving psychotherapy, to the hospitalized subjects.

The one exception, Subject D, a "normal" subject, received the second lowest rating of Judge 1 and the lowest

rating of Judge 2. In 3 out of 4 comparisons where Subject D was a member of the pair, the judges agreed in picking Subject D as being the lesser adjusted member of the pair. In the case of the one disagreement, Subject D was picked as being healthier than a hospitalized subject. Independent analysis of the TAT stories of this subject by two judges not used in the study showed that even though this subject was functioning outside an institution, he was severely maladjusted. Attempts to obtain further diagnostic material on this subject were not successful.

The questionnaire given to the judges to be completed after the task indicated that they used the criteria outlined in the instructions more than they used their own subjective criteria, but that they felt the criteria agreed with their own conception of emotional health. One judge reported more use of the TAT criteria than the other criteria while one judge reported using both equally. Both judges stated that the criteria helped them in making their judgments. Neither felt that judging adjustments, which is somewhat contrary to the usual type of judgment involved in clinical judgment studies, interfered with their judgment although both felt that this emphasis was different.

DISCUSSION

The Technique

Intra-judge reliability, R_{ss}, was .73 and .93, which indicates that emotional health was reliably scaled on a unidimensional ratio scale, since the discrepancy variance includes failures of the theoretical model such as departures from unidimensionality or lack of ratio scale properties, as well as errors of judgment, fatigue and carelessness. It is clear that this technique of scaling makes clinical judgment a quantitative measuring device as least as reliable as most objective tests. Moreover, the correlation between the judges of .87 is certainly high enough to consider the two judges parallel forms of the same test.

Even though the method used in this study bears some similarity to the paired comparison technique of Thurstone (1926, 1927a, 1927b), it has at least two advantages which seem to make it more desirable as a method to be used in clinical judgment studies. Thurstone's method, as Guilford points out (1929, 1931, 1954), requires a great deal of computation. Derivation of scale values by the technique used in this study requires much less computation. Moreover, Thurstone's method, as well as Guilford's modification of it (1928, 1931), requires that stimuli be judged many

times, either by one judge, judging many times, or by many judges, each judging one time.

Disregarding the use of many judgments produced by one judge, which are frequently found to be in error, the use of many judges presents a hinderance to the use of either Thurstone's or Guilford's method with clinical material. Finding large numbers of qualified judges to participate at one time in a research project is almost impossible. The method used in this study overcomes this difficulty since scale values can be obtained from single judgments of paired stimuli by as few as one judge. Naturally, in practice, more than one judge would be used.

The Attribute

Most often clinical training consists of focusing on pathology so that the clinician is set to see signs of psychopathology and to make his judgments accordingly. Insofar as judgments are made on the basis of signs, there is the risk of relying on indications which have been shown to result in judgments of low reliability (Elikins, 1958).

The use of psychopathology in clinical judgment studies, as an attribute to be judged, may not be optimal since it is impossible to establish a base line of illness. The use of emotional health as the attribute to be judged avoids this difficulty since it is difficult to conceive

of anyone as being completely without health. Thus, emotional health has at least a conceivable point of origin.

While it is impossible to say what bounds or limits there are to emotional health, it is fairly safe to assume that no one ever achieves his fullest potential. Moreover, while any one aspect of emotional health may be taken as indicating the presence or absence of the attribute, only a human judge is able to evaluate simultaneously all the interrelationships of the various components and produce a single judgment.

Anchors and the Amount of Material

Four naive judges were used in an exploratory study where the task was to judge emotional disturbance using TAT stories in pairs. Scale values and reliability coefficients for the exploratory study are shown in the Appendix. Scale values for the naive judges had a spread of 5.1 units, while scale values for the two experienced judges had a spread of 2.8 units. Whether or not the example TAT stores representing both extremes of pathology which were used in instructing the experienced judges, but not the naive judges, acted as anchors for the experienced judges, is a matter of speculation since no specific test for such effects were included in this study. However, as Hunt (1941) points out, the use of anchors serves to extend the rating scale and results in a greater tendency for judgments to be nearer

the middle of the scale. It is possible that anchor effects were operating in the judgments of the experienced judges and resulted in less spread of the scale values. If this were so, it would indicate the importance of having supplied anchors in clinical judgment studies for both ends of the continuum rather than leaving it up to the judges to develop their own anchors as is the case when anchors are not supplied.

The use of 20 TAT stories in each of the six sets represents a large amount of material for each judge to process. The finding that such reliable judgments could be made by experienced judges contradicts the findings of several studies, but may be explained by the finding of Newton (1954) that reliable judgments could be made using large amounts of material if the judges were allowed time for exhaustive analysis of the material. The judges in this study made their judgments over a two-week period of time at their leisure.

The Use of the Method

One advantage to ratio scale values is, as Torgerson (1958) points out, that the difference between the ratios of the scale values can be interpreted as reflecting the differences of the stimuli, as well as transitivity, so that if A is judged greater than B, and B is judged greater than C, then A can be assumed to be greater than C and the differences between the scale values can be interpreted as

reflecting the differences between the properties being judged.

Given any complex entity composed of inter-related, identifiable aspects, it would seem to be possible to use this measurement technique in a series of judgment studies where each identifiable aspect was isolated and used as a single criteria for the attribute being judged. Thus scale values derived for the attribute being judged on the basis of different aspects of the attribute could be compared and the relative contribution of each to the formation of judgments about the attribute could be evaluated according to the property of ratio scale values mentioned above. That is, if judgments of A, using aspect Z, resulted in scale values of 2.00, and judgments of A using aspect Y, resulted in scale values of 1.00, it would be reasonable to assume that judgments of A were affected more by Z than Y when both were used as criteria. Obviously, the reliability of judgments made on the basis of each aspect could be determined in order to see which aspect afforded the greater reliability.

While this would, in effect, result in a "factoring" out of the dimensions along which judgments are made, such a "factoring" would be more closely tied to the subjective use of the dimensions than would seem to result when formal methods of factor analysis are used. In this way, clinical judgment research would come closer to studying the actual process of forming judgments than has resulted

in research which has relied primarily on correlational analysis.

Since emotional health is, as Johoda (1959), Scott (1958), and Epstein (1958) point out, comprised of many components, it can be regarded as a multidimensional attribute. Since this multidimensional concept was scaled on a unidimensional scale, it is likely that other attributes as complex as emotional health may also be scaled, so that it seems feasible to use this measuring technique to compare whole entities rather as well as parts of one.

SUMMARY

The purpose of this paper was to determine if judgments of emotional health could be measured using Estavan's modified paired comparison method, and scale values derived for the stimuli judged.

TAT stories were judged in $\frac{n(n-1)}{2}$ pairs by two experienced clinical graduate students for emotional health. Following Estavan's method, scale values were derived for each stimulus judged. Inter-judge reliability was found to be .87. Intra-judge reliability was found to be .79 for one judge and .93 for the others.

The method of developing scale values as used in this study bears some strong resemblance to Thurstone's Case V method, but has definite advantages over the Case V method.

BIBLIOGRAPHY

Albee, G. W. and Hamlin, R. M. An investigation of the reliability and validity of judgments of adjustment inferred from drawings. J. clin. Psychol., 1949, 5, 389-392.

Allison, Roger, Jr., Korner, I. N., and Zwanziger, M. D. Clinical judgments and objective measures. J. Psychol., 1964. 57, 451-456.

Arnhoff, F. N. Some factors influencing the unreliability of clinical judgment. J. clin. Psychol., 1954, 10, 272-275.

Bialick, J. and Hamlin, R. M. The clinician as judge: Details of procedure in judging projective material. J. consult. Psychol., 1954, 230-242.

Block, W. E. A study of meaning set in the judgment of clinical test data. J. clin. Psychol., 1962, 18, 511-512.

Block, W. E. Adaptation effects in clinical judgment of projective test data. J. clin. Psychol., 1964, 20, 448-454.

Borke, H. and Fiske, D. W. Factors influencing the prediction of behavior from a diagnostic interview. J. consult. Psychol., 1957, 21, 78-80.

Campbell, D. T., Hunt, W. A., and Lewis, N. A. The effects of assimilation and contrast in judgments of clinical materials. Amer. J. Psychol., 1957, 70, 347-360.

Cummings, S. T. The clinician as judge: Judgment of adjustment from Rorschach single card performance. J. consult Psychol., 1954, 18, 243-247.

Elkins, E. Diagnostic validity of the Ames "danger signals". J. consult Psychol., 1958, 22, 281-287.

Epstein, N. B. Concepts of normality or evaluation of emotional health. Behv. Sci., 1958, 3, 335-343.

Giedt, F. H. Comparison of visual, content and auditory cues in interviewing. J. consult. Psychol., 1955, 19, 407-416.

Goldfarb, A. Reliability of diagnostic judgments made by psychologists. J. clin. Psychol., 1959, 15, 392-396.

Grant, M. Q., Ives, V., and Ranzoni, J. H. Reliability and validity of judged ratings of adjustment on the Rorschach. Psychol. Monogr., 1952, 66, No. 2 (Whole Number 334).

Griggs, A. E. Experience of clinicians and speech characteristics and statements of clients as variables in clinical judgments. J. consult. Psychol., 1958, 22, 315-319.

Gross, C. F. Intra-judge consistency in ratings of heterogeneous persons. J. abnorm. soc. Psychol., 1961, 62, 605-610.

Grosz, H., and Grossman, K. The source of observer variation and bias in clinical judgment. I: The item of psychiatric history. Journal of Nerv. and Ment. Dis., 1964, 138, 105-113.

Guilford, J. P. The method of paired comparisons as a psychometric method. Psychol. Rev., 1928, 35, 494-506.

Guilford, J. P. Some empirical tests of the method of paired comparisons. J. gen. Psychol., 1931, 5, 64-76.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Gulliksen, H. and Tukey, J. W. Reliability for the law of comparative judgment. Psychometrika, 1958, 23, 95-110.

Gunderson, E. K. E. Determinants of reliability in personality ratings. J. clin. Psychol., 1965, 21, 164-169.

Hamlin, R. M. The clinician as judge: Implications of a series of studies. J. consult. Psychol., 1954, 18, 233-238.

Holsopple, J. Q., and Phelan, J. G. The skills of clinicians in analysis of projective tests. J. clin. Psychol., 1954, 10, 307-320.

Horowitz, M. J. A study of clinicians judgments from projective test protocols. J. consult. Psychol., 1962, 26, 251-256.

Hunt, W. A. Anchoring effects in judgments. Amer. J. Psychol., 1941, 44, 395-403.

Hunt, W. A., Arnhoff, F., and Cotton, J. Reliability, chance and fantasy in interjudge agreement among clinicians. J. clin. Psychol., 1954, 10, 296-299.

Hunt, W. A. and Jones, N. F. Clinical judgments of some aspects of schizophrenic thinking. J. clin. Psychol., 1958a, 14, 235-239.

Hunt, W. A. and Jones, N. F. The reliability of clinical judgments of asocial tendency. J. clin. Psychol., 1958b, 14, 233-235.

Hunt, W. A. and Jones, N. F. The experimental investigation of clinical judgment. In A. J. Bachrach (Ed.), Experimental foundations of clinical psychology. New York: Basic Books, 1962.

Hunt, W. A., Jones, N. F., and Hunt, E. B. Reliability of clinical judgments as a function of clinical experience. J. clin. Psychol., 1957, 13, 377-378.

Hunt, W. A., Schwartz, M. L., and Walker, R. E. Reliability of clinical judgments as a function of range of pathology. J. abnorm. Psychol., 1965, 70, 32-33.

Hunt, W. A. and Walker, R. E. A comparison of global and specific clinical judgments across several diagnostic categories. J. clin. Psychol., 1962, 18, 188-194.

Jackson, M. A. The effects offrequency, extremeness, consistency and order of the stimulus on clinical judgments. Diser. Abst., 1963, 24, 1244.

Jahoda, M. Current conceptions of positive mental health. New York: Basic Books, 1958.

Johnson, D. M. The psychology of thought and judgment. New York: Harper and Brothers, 1955.

Jones, N. F. Context effect in judgment as a function of experience. J. clin. Psychol., 1957, 13, 379-382.

Jones, N. F. The validity of clinical judgments of schizophrenic pathology on verbal responses to intelligence test items. J. clinc. Psychol., 1959, 396-400.

King, G. F., Ehrmann, J. C. and Johnson, D. M. Experimental analysis of the reliability of observations of social behavior. J. soc. Psychol., 1952, 35, 151-160.

Kostlan, A. A method for the empirical study of psychodiagnosis. J. consult. Psychol., 1954, 18, 83-88.

Lee, J. C. and Tucker, B. An investigation of clinical judgment: A study in method. J. abnorm. soc. Psychol., 1962, 64, 272-280.

Levine, H. The influence of fullness of interview on the reliability, discriminability and the validity of interview judgments. J. consult. Psychol., 1954, 18, 303-306.

Levy, L. H. Context effects in social perception. J. abnorm. soc. Psychol., 1960, 61, 295-297.

Little, K. B., and Shneidman, E. S. Congruencies among interpretations of psychological tests and anamnestic data. Psychol. Monogr., 1959, 73, No. 42 (Whole No. 476).

Luborsky, L. The patient's personality and psychotherapeutic change. In Strupp, H. and Luborsky, L. Research in Psychotherapy. Washington: American Psychological Assn., 1962.

Luft, J. Implicit hypotheses and clinical prediction. J. abnorm. soc. Psychol., 1950, 45, 756-760.

Luft, J. Differences in prediction based on hearing vs. reading verbatim clinical interviews. J. consult. Psychol., 1951, 15, 115-119.

Mancuso, C. J. The role of influences of sociological variables in clinical judgment. Diser. Abst., 1961, 21, 2785.

Martin, H. T., Jr. The nature of clinical judgments. Diser. Abst., 1958, 18, 310.

Mehlman, G. The reliability of psychiatric diagnosis. J. abnorm. soc. Psychol., 1952, 47, 577-578.

Miller, H. and Bieri, J. An informational analysis of clinical judgment. J. abnorm. soc. Psychol., 1963, 67, 317-325.

Miller, N. and Campbell, D. T. Recency and primacy in persuasion as a function of the timing of speeches and measurement. J. abnorm. soc. Psychol., 1959, 59, 1-9.

Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. Psychometrika, 1951, 16, 3-9.

Newton, R. L. The clinician as judge; total Rorschach and clinical case material. J. consult. Psychol., 1954, 18, 248-250.

Pasamanick, B., Dinitz, S., and Lefton, M. Psychiatric orientation and its relation to diagnosis and treatment in a mental hospital. Amer. J. Psychiat., 1959, 116, 127-132.

Phelan, J. G. The subjective feeling of certainty of diagnostic judgment of clinical psychologists. J. clin. Psychol., 1960, 16, 101-104.

Phelan, J. G. Rationale employed by clinical psychologists in diagnostic judgments. J. clin. Psychol., 1964, 20, 454-458.

Phelan, J. G. Use of matching methods in measuring reliability in individuals. Psychol. Reps., 1965, 16, 490-496.

Powers, W. T. and Hamlin, R. M. The validity, basis and process of clinical judgment using a limited amount of projective test data. J. proj. Tech., 1957, 21, 286-293.

Raines, G. N. and Rohrer, J. H. Individual differences in clinical judgment. Amer. J. Psychiat., 1955, 110, 721-725.

Scott, W. A. Research definitions of mental health and mental illness. Psychol. Bull., 1958, 55, 29-45.

Sines, L. K. The relative contribution of four kinds of data to accuracy in personality assessment. J. consult. Psychol., 1959, 23, 483-492.

Soskin, W. F. Frames of reference in personality assessment. J. clin. Psychol., 1954, 10, 107-114.

Soskin, W. F. Influence of four types of data on diagnostic conceptualization in psychological testing. J. abnorm. soc. Psychol., 1959, 58, 69-78.

Thorne, F. C. Clinical Judgment. Brandon, Vermont: Journal of Clinical Psychology, 1961.

Thurstone, L. L. The method of paired comparisons. J. abnorm. soc. Psychol., 1926, 21, 384-400.

Thurstone, L. L. A law of comparative judgment. Psychol. Rev., 1927a, 34, 273-286.

Thurstone, L. L. Psychophysical analysis. Amer. J. Psychol., 1927b, 38, 368-389.

Torgerson, W. Theory and Method of Scaling. New York: Wiley, 1958.

Weitman, M. Some variables related to bias in clinical judgment. J. clin. Psychol., 1962, 504-506.

Wessman, A. E. and Ricks, D. F. Mood and Personality. New York: Holt, Rinehart and Winston, 1966.

APPENDICES

APPENDIX A

Table A. Classification and personal data of subjects giving TAT protocols.

NORMALS	SUBJECT A	SUBJECT D
Age	19	19
Sex	Male	Male
Education	Sophomore	Sophomore
Mother	Living	Living
Age	45	40
Occupation	Office Worker	Housewife
Father	Living	Living
Age	44	47
Occupation	Office Manager	CPA
Siblings	Three	None
Sex, Age	Male	
Occupation	21 College	
	12 High School	
	5	
HOSPITALIZED	**SUBJECT B**	**SUBJECT E**
Age	20	26
Sex	Male	Male
Education	High School	High School
Mother	Living	Living
Age	39	47
Occupation	Housewife	Housewife
Father	Living	Living
Age	39	51
Occupation	Mechanic	Post Office Employee
Siblings	Two	One
Sex, Age	Male	Female
Occupation	16 High School	Housewife
	10 Grammar School	
Diagnosis	Schizophrenic	Schizophrenic
Length of Hospitalization	Three months	Four months
COUNSELING CENTER	**SUBJECT C**	**SUBJECT F**
Age	20	20
Sex	Male	Male
Education	Junior	Junior
Mother	Living	Living
Age	42	49
Occupation	Housewife	Housewife
Father	Living	Living
Age	40	43
Occupation	TV Repair	Fireman
Siblings	Two	One
Sex, Age	1 Male, 17, High	Female, 21 Office
Occupations	School; 1 Female, 11	Work

APPENDIX B

Table B. Order at TAT card presentation to subjects.

Subjects					
Normals		Counseling Center		Hospitalized	
A	D	B	E	C	F
1	1	1	1	1	1
2	2	2	2	2	2
11	6 BM	4	11	11	11
10	10	6 BM	9 BM	4	9 BM
17 BM	11	5	18 BM	13 MF	17 BM
18 BM	9 BM	9 BM	17 BM	19	12 M
12 M	12 M	7 BM	12 M	7 BM	18 BM
6 BM	18 BM	8 BM	14	17 BM	6 BM
14	17 BM	3 BM	3 BM	18 BM	20
9 BM	4	10	10	9 BM	14
3 BM	13 MF	15	6 BM	12 M	3 BM
13 MF	5	20	5	14	5
7 BM	15	19	19	3 BM	4
8 BM	3 BM	18 BM	15	10	15
20	14	14	20	6 BM	13 MF
15	19	13 MF	7 BM	5	19
5	20	12 M	8 BM	15	10
19	7 BM	17 BM	4	20	7 BM
4	8 BM	11	13 MF	8 BM	8 BM
16	16	16	16	16	16

APPENDIX C

INSTRUCTIONS TO JUDGES

INTRODUCTION

You are being asked to act as a judge in research which I am doing for my Master's Thesis. The thesis is concerned with whether or not clinical judgments can be quantified and compared over judges. The materials to be judged are 6 sets of TAT stories, with 20 stories to a set. The sets are to be judged in pairs. The judgments you will be making are concerned with the amount of emotional health one person has when he is compared with another person.

I have outlined below the criteria I wish you to use in making your judgments as well as a method to use.

CRITERIA

As you read the TAT stories, keep in mind the following criteria of emotional health. You will find that the criteria are divided into two sets. One set describes some components which we believe are involved in emotional health, while the second set describes indications or signs of emotional health as it might appear specifically in TAT stories.

In making your judgments, use both sets of criteria, but remember that they are not absolute. In the end, rely upon your own subjective, clinical judgment, and let these criteria only be guides to that judgment.

Components of Emotional Health

Ability to take care of self

Ability to work

Sexual adjustment

Social adjustment

Absence of hallucinations, bizarre delusions, gross distortion of reality, lack of passivity.

Degree of freedom from anxiety and depression, degree of diffuse hostility.

Amount of affect, owning of feelings.

Variety and spontaneity of affect

Satisfaction with life and with self, absence of deficiency motivation, i.e., making up for lost love

Achievement of capabilities, mastery of environment

Benign rather than malignant effect on others

Indications of Emotional Health in TAT Stories

The protocols should be longer

There should be more affect, and more varied affect

There should be less stereotyped, and more varied materials, e.g., the TAT stories should vary more from card to card indicating an ability to deal with differing aspects of the world

There should be more benign fantasies and more helping parent figures.

There should be good reality testing

Problems should be directly represented

There will be indications of confidence

METHOD

You will be judging the protocols in pairs. When grouped together, there are 36 possible pairs. When you eliminate pairs because of duplications, such as, (AB, BA), and (BF, FB), and self-comparisons, such as, (AA) and (BB), you are left with only 15 pairs. We are concerned with these 15 pairs. I have listed them on the last sheet of these instructions.

Take one pair at a time, according to the order in which I have listed them. Read each protocol of each pair as you judge the pair. The first protocol to be read is the first one listed in the pair. For example, of the pair (A,F), read protocol A first, then read F; of the pair (D,E), read protocol D first, then read E.

After you have completed a pair, judge, according to the criteria outlined above, which protocol seems to represent the person who has the most psychological health of the pair. At the time you are judging, you may wish to reread parts of one or both protocols. You may do so. Do not, however, compare them as you are first reading them through.

Take a sheet of the paper on which I have drawn a line. Label the sheet with the letter representing which member of the pair you have judged to be healthier. Let the line represent the total amount of health the healthier member has. In comparison to this amount, mark off some point on the line which indicates how much of this health

the second member of the pair has. For example's sake, let's suppose you are considering the hypothetical pair (Z,X), and you think that Z is the healthier member of the pair. Label the sheet Z. Let the line equal the total amount of health Z has. Suppose you think that in comparison to Z, X has about half as much health. Place a mark in the middle of the line. Continue on for each of the other pairs. I have marked the sheets so that you will be able to identify easily the pairs as you are marking the sheets. After you are finished judging, I would appreciate it if you would answer the questionnaire I have included with these instructions.

THE PROTOCOLS

The protocols were obtained by administering 20 TAT cards to six subjects. All subjects were given the same card, and asked to make up stories to them. Their stories were first recorded on tape and then transcribed to paper.

The stories are as near as possible to verbatim. In transcribing the stories, no effort was made to altar the stories in any way, so that the story as told could be judged. Pauses, when the subject seemed to be groping for words, have been indicated by a series of dots (.....). The number of dots does <u>not</u> indicate the length of the pause. Long silences, when the subject seemed to be searching for ideas are indicated with the words (Long Silence). Comments or questions made by the tester during the session have been enclosed in brackets so as to distinguish them from the

story proper.

On the face sheet of each protocol, you will find information about the subject's age, sex and number of siblings in the family. This should help you in your judgments.

APPENDIX D

Table C. The order in which protocols were presented to the judges in pairs for comparison and judgment.

(A, F)
(D, E)
(B, F)
(C, A)
(E, F)
(C, D)
(B, E)
(A, D)
(C, E)
(D, F)
(B, C)
(A, E)
(B, D)
(C, F)
(A, B)

APPENDIX E

Table D. Member of each pair judged to be healthier by two judges.

Pair	Judged Healthier	
	Judge 1	Judge 2
(A, F)	(A)	(A)
(D, E)	(D)	(E)
(B, F)	(F)	(F)
(C, A)	(A)	(A)
(E, F)	(F)	(F)
(C, D)	(C)	(C)
(B, E)	(B)	(B)
(A, D)	(A)	(A)
(C, E)	(C)	(C)
(D, F)	(F)	(F)
(B, C)	(C)	(C)
(A, E)	(A)	(A)
(B, D)	(B)	(B)
(C, F)	(C)	(C)
(A, B)	(A)	(A)

APPENDIX F

Table E. Scale values for 6 stimuli judged by 4 naive judges using an attribute of emotional disturbance.

Stimulus	1	2	3	4
A	3.147	.3910	.7100	1.311
B	2.307	4.124	1.539	.7573
C	.8902	2.036	3.959	1.993
D	.2650	.3062	.4173	.3268
E	.1865	.1875	.2873	.3506
F	3.131	5.269	1.927	4.313

APPENDIX G

Table F. Intra-judge reliability for 4 naive judges.

Judge 1	-.24
Judge 2	+.80
Judge 3	-.22
Judge 4	-.16

Table G. Inter-judge reliabilities for 4 Naive judges (Pearson r).

Judges	r
(1,2)	.58
(1,3)	.09
(1,4)	.60
(2,3)	.45
(2,4)	.72
(3,4)	.64

Table H. Inter-judge reliabilities for 4 naive and 2 experienced judges (Pearson r).

Judges	r
(1,A)	.63
(1,B)	.22
(2,A)	-.08
(2,B)	.21
(3,A)	.21
(3,B)	.46
(4,A)	.20
(4,B)	.18

APPENDIX H

QUESTIONNAIRE

The following questions were asked of the experienced judges after completion of their task.

1. In judging the stories did you rely more on your own clinical judgment or upon the criteria outlined in the study?
2. Did you use the criteria outlined to make your judgments?
3. In making your judgments, which criteria did you use most, if either?
4. Did the criteria help or hinder in any way your making judgments?
5. Did the emphasis on emotional health rather than sickness seem different to you or interfer with your judgment?
6. Did the criteria outlined in the study clash with your own conception of emotional health?
7. Which of the criteria outlined in the study did you find to be the most helpful?
8. Which of the criteria outlined in the study did you find to be the least helpful?
9. Was there enough material in the stories so you could judge on the basis of the criteria?
10. Could you use the criteria with TAT stories?
11. Do you think this kind of judgment is made more accurate due to the comparison of one person with another?