THE RELIABILITY AND INTERNAL CONSISTENCY OF THE THEMATIC APPERCEPTION TEST

by

MARILYN EPSTEIN

B.A., Brooklyn College, 1947

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Arts

in the Department of

Psychology

We accept this thesis as conforming to the standard required from candidates for the degree of MASTER OF ARTS

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1964

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Psychology

The University of British Columbia,
Vancouver 8, Canada

Date April 8, 1964

ABSTRACT

The purpose of this study was to investigate the repeat reliability and internal consistency under short-term conditions of several indices of agression and anxiety as measured by the TAT.

In view of the variations in the results reported in the few studies concerned with this problem, a specificity hypothesis was suggested. This hypothesis states that no general evaluation can be made of the temporal and internal stability of the TAT. Such statements probably only have meaning in terms of specific variables. The variables employed in the present study were aggression and anxiety and the results should not be generalized beyond these variables.

One group of subjects was given standard TAT instructions at two successive administrations, while a second group was asked to tell a different story to each card. This procedure was designed to control and study the influence of memory effects. It was found that memory effects are very strong, and where the instructions interfere with their operation, repeat reliability coefficients are very low.

The TAT cards included two high, two medium and two low aggressive content cards, as determined by a panel of judges and from previous research. The purpose of this part of the study was to determine if the reliability of the test varies with the level of card ambiguity for a given drive. The results did not support the hypothesis that responses to stimuli which are unambiguous for a given drive are more likely to be stable over time than responses made to a relatively ambiguous stimulus.

The internal consistency was evaluated by correlating the scores obtained on the first session by all subjects in terms of the level of ambiguity. These correlations were quite low, indicating the need for caution in using an additive treatment of scores from different TAT cards.

ACKNOWLEDGEMENT

The writer wishes to give special acknowledgement to Dr. D. T. Kenny under whose supervision this thesis was carried out. Not only did the problem originate with him, but the subsequent work would never have been brought to fruition without his continuous encouragement and advice.

Thanks are also due to Dr. D. C. G. MacKay for a critical reading of the thesis; to Mrs. Joyce Treit, who acted as independent judge for several aspects of the work; and to the many students who contributed generously of their time to act as subjects.

CONTENTS

LIST OF TABLES

CHAPTER I

THE PROBLEM AND ITS BACKGROUND

While research concerned with the Thematic Apperception Test (TAT) has been extensive, very little evidence exists on either the internal consistency or repeat reliability of the test.

One likely reason for this lack is the claim of some psychologists that test-retest reliability is not to be expected in projective instruments because it is almost impossible to attain the same motivational situation in successive administrations or, as McClelland says, "to put the subject back in the condition he was in before he made the first response" (1958, p. 20). Tomkins has compared this situation to the difficulty of trying to measure the reliability of a response to a joke; if a joke told twice in succession to the same person does not produce the same response both times, no inference about the reliability of the response can be made (Lesser, 1961). Kagan (1960) suggests that the question of reliability probably only has meaning with respect to specific variables scored from specific stimuli and considers this similar to a specific blood test, where it is not expected that the test will be a reliable index of all compounds in the blood. This comparison is probably not valid in view of actual clinical practice with the TAT, where it is commonly used to give a global picture of personality. However, for research purposes this appears to be a fruitful approach.

The increasing concern with theoretically-oriented research on the TAT and the growing recognition of the inadequacy of the purely empirical approach (Lindzey, 1958) is another reason for the lack of interest in investigations of the psychometric aspects of the TAT. The time and effort involved in collecting data for reliability studies has obviously seemed incommensurate with the importance of the strictly empirical results to be derived from such studies.

Nevertheless, a reasonable case can be made for studies on the temporal and internal stability of TAT measurement. Reliability of measurement has been the minimal requirement of all scientific data. In applying this criterion to the TAT it should be noted that the summation of scores for similar story-events from different cards assumes that the scores are "tapping" similar psychological processes. If the scores do co-vary together, then the TAT should possess internal consistency. It has been suggested (Jensen, 1959) that any additive treatment of TAT variables is similar to adding together pounds, gallons and inches. To refute this argument requires many more data on the internal stability of the test than are currently available.

While certain procedural problems intrude into studies of the test-retest reliability of the TAT, it would nevertheless seem desirable to have some consistency of measurement in fantasy assessment. Lindzey and Herman (1955), in one of the few studies on the reliability of the TAT, concluded that traditional questions about

reliability should be asked about the TAT since the "...answers to such questions will prove necessary eventually for a full understanding of the clinical function of these instruments. In view of this, even very fragmentary findings, if they offer any possibility of cumulating with the results of other studies, are highly desirable and to be encouraged" (p. 41-42). It seems clear, therefore, that a crucial task now facing the TAT researcher is the collection of reliability data.

The main aim of the present study is to determine the internal consistency and repeat reliability of the TAT under short-term conditions. Repeat reliability studies of the TAT have generally used a relatively long interval between successive administrations of the test and were primarily concerned with the long-range stability of various fantasy contents. From a psychometric point of view this does not provide satisfactory evidence for test-retest reliability.

The primary focus in the present investigation is on the test-retest reliability and internal consistency of various indices of aggression and anxiety. One of the purposes of the study is to investigate the effects of instructions at the second administration of the test. The general procedure in this research requires one group of subjects to respond to the cards the second time under identical first administration instructions and another group to tell their stories to the cards under instructions to make up a new story.

Another purpose of the present study is to examine the effects

of drive-structure of TAT cards on the temporal stability of aggression and anxiety. In conformity with Kagan's hypothesis (1955) it was predicted that temporal stability of fantasy variables would be a direct positive function of the drive-structure of TAT cards.

The third purpose of the present research is to investigate the internal consistency of aggression and anxiety indices. In general, it seems reasonable to expect fantasy indices of aggression and anxiety to covary positively within a given set of TAT cards.

CHAPTER II

REVIEW OF RELATED RESEARCH

In view of the vast quantity of research on the TAT the review of the literature will be limited to those studies bearing directly on the problem of reliability and internal consistency of this test.

The TAT variable which has been studied most extensively is the achievement motive. In an unpublished study by E. L. Lowell, described by McClelland et al. (1953), two equivalent forms of Atkinson's achievement pictures were administered to the same group of forty male college students with an interval of one week between measures, and a product-moment correlation of .22 was reported. The authors draw attention to the difficulty of being certain that the periods of stimulation immediately preceding the two measures were equivalent for each subject on the two administrations and this is offered as a possible explanation of the low reliability. Atkinson (1950) found the product-moment correlation to be .64 between these same two equivalent forms when they were administered at the same time. These pictures include three TAT cards (Cards 1, 7 BM and 8 BM). The other pictures are from other sources.

Haber and Alpert (1958) report a test-retest reliability correlation of .54 for achievement with a comparable set of pictures and an equivalent experimenter used in the second session.

The interval between administrations was three weeks. The same study also analyses the responses to the pictures in terms of ambiguity of the stimulus for the achievement motive. The repeat reliability of the low-cue pictures taken separately is .36, while for the high-cue pictures it is .59. The correlation between the high-cue pictures and the low-cue pictures is .57. These findings would seem to lend some support to Kagan's hypothesis (1955) that content categories reported to stimuli ambiguous for that content are less likely to be stable than those produced by stimuli which suggest that content.

Both McClelland and Atkinson postulate a "set for response variability" to account for low test-retest reliability and suggest that thematic apperceptive responses may show cyclical alternation over three successive administrations.

As a by-product of research on the relation between TAT performance and self-ratings, Child et al. (1956) reported reliability coefficients of internal consistency ranging from -.07 to +.34, with a mean of .13 on ten major Murray TAT variables. The test was administered on a group basis. The authors point out that these correlations are far lower than the reliability of the self-rating questionnaires they used to measure the same variables. These results have prompted one reviewer to state that "any scoring system based on the addition of themes elicited by various pictures is fallacious. A theme on one card is not sufficiently correlated

with the same theme on another card to justify an additive treatment of TAT variables" (Jensen, 1959, p. 311).

Auld et al. (1955) report a test-retest reliability coefficient of .13 and a rank order correlation of .10 for sexual motivation using Guttman derived scales for TAT scoring. The above coefficients are for eighteen subjects who were enclosed in a sealed submarine for over a month. The same study reports an unsuccessful attempt to construct a reliable scale of the same type to measure aggression.

Tomkins (1947) concludes that repeat reliability is a function of the time interval between successive administrations, that is, as the time interval increases the reliability declines, except where the personality of the individual is extremely stable. He reports reliability coefficients of .80 for subjects retested after two months, .60 when there was a six-month interval, and .50 when there was ten months between administrations. No further details are given on this study and therefore it is difficult to evaluate these results. Tomkins states that the protocols were analysed according to "Murray's quantitative need-press scheme" but gives no further information about the variables employed.

In a study on reliability and situational validity Lindzey and Herman (1955) report split-half reliability coefficients ranging from .12 to .45 with eight cards on six variables. Each story was scored on a five-point rating scale for each of the variables. The authors emphasize that these findings apply only to the story rating

method of quantifying protocols and suggest using different units of analysis.

In the same paper Lindzey and Herman describe an investigation of repeat reliability where subjects were asked, after a two month interval, to tell a story different from the first story they had told to each card. There were twenty subjects involved who told stories to four TAT cards and the stories were scored for seventeen variables. The correlations ranged from .00 to .94, with consistently high standard errors (.17 to .72).

In a very recent study on intraindividual consistency of TAT stories when the test is administered under several different conditions, Wylie et al. (1963) report reliability coefficients for Aggression and Dependency. They used twelve TAT cards, divided into two sets of six each, which they judged to be equal in "pull" for these two variables. To two of their groups the test was administered twice with typical instructions and a one-week interval between testing sessions. The reliability coefficients were .43 for Aggression and .38 for Dependency. They concluded that the level of intraindividual consistency is too low for reliable individual diagnosis.

Table 1 shows a summary of the research evidence related to TAT reliability and internal consistency. Inspection of this table reveals that researchers have paid scant attention to reliability studies. Examination of the data in Table 1 shows that there is

very little evidence to support the belief of either the temporal or internal stability of the TAT. However, in terms of the variation in the few correlations which have been reported, a specificity hypothesis, as suggested by Kagan (1960), may be assumed. This hypothesis would assert that no general statement may be made about TAT reliability or internal consistency. If this assumption is valid, then these two psychometric aspects of the TAT would vary as a consequence of the variable being scored. On this basis it would be expected that repeat reliability and internal consistency would differ for various TAT scoring scales. If this hypothesis is reasonable, then it would not be sensible to ask, in general, what the reliability of the TAT is. Rather, one should ask, for example, what is the reliability of TAT aggression, as scored by specific criteria.

The present study is designed to go a little way toward filling the research gap on two major TAT variables, namely, those of aggression and anxiety. If the specificity hypothesis is correct, the evidence obtained from this research could not be generalized beyond these two variables.

TABLE 1

SUMMARY OF REPEAT RELIABILITY AND INTERNAL CONSISTENCY CORRELATIONS AS REPORTED IN THE LITERATURE

REPEAT RELIABILITY

	N	Method	Time Interval	Variable	Correlation
Lowell	40	Equivalent form	One Week	Achievement	.22
Haber and Alpert	26	Equivalent form	Three Weeks	Achievement	.54##
Auld et al.	18	Same form	One Month	Sex	.13
x Tomkins	15	Same form	Two Months	"Murray's quantitative	.80
	15		Six Months	need-press scheme"	.60
	15		Ten Months		.50
Lindzey and Herman	20	Same form	Two months	n Abasement	.32
				n Affiliation	.00
				n Autonomy	.49 #
				n Cognizance	.49 #
				n Counteractive Achievement	.67 ##
				n Recognition	.94 ##
				p Dominance	.45 #
				p Rejection	.66 ##
				Hero Assists Others	.50 #
				Hero Assisted by Others	.67 ##
				Story Outcomes	.50 #
				Achievement of Goals	.07
				Failure to Achieve Goals	.86 ##
				Tension Relief Words	.82 ##
				Food Words (goal)	.60 ##
				Food Words (instrumental)	.28
				Total Food Words	.30

(Table continued on next page)

INTERNAL CONSISTENCY

	N	Variable	Correlation
Atkinson	40	Achievement	.64 ##
Lindzey and Herman	148	n Achievement	.19 #
		n Aggression	.29 ##
		n Sex	.45 ##
		n Abasement	.28 ##
		n Nurturance	.12
		Narcissism	.20 #
Child et al.	183	Achievement	+.27 ##
		Aggression	+.34 ##
		Autonomy	+.21 ##
		Deference	+.30 ##
		Dominance	+.10
		Isolation	-.02
		Nurturance	-.07
		Responsibility	-.06
		Sociability	+.10
		Succorance	+.12
Wylie et al.	24	Aggression	.43 #
		Dependency	.38

\# P <.05
P <.01

x It is impossible to evaluate the significance of the correlations reported in this study because no details are given on the methods employed.

CHAPTER III

EXPERIMENTAL METHOD

Selection of cards

Seventeen volunteers from a senior course in psychology at the University of British Columbia were asked to rank order nine TAT pictures (Cards 1, 2, 3BM, 4, 6BM, 11, 14, 18BM, 18GF) in terms of the amount of hostility expressed in them. (See Appendix I for instructions.) These cards were selected because they are thought to represent a range of aggression in terms of their card pull.

There were four females and thirteen males in the group, and their ages ranged from twenty to forty-two, with a mean of 23.82 and a standard deviation of 4.88. Table 2 shows the frequency distributions for their ratings.

It can be seen from Table 2 that some cards are more consistently rank ordered than others, as can be readily assessed by variation in the number of subjects assigning different ranks to the same picture. For example, card 18GF is generally regarded by these subjects as aggressive, whereas for card 6BM the subjects do not especially agree amongst themselves as to its aggressiveness.

On the basis of these judgements cards 1 and 14 were selected as low-aggressive, cards 3BM and 11 as moderately aggressive, and cards 18BM and 18GF as highly aggressive. Previous research lends strong support for the present classification of TAT cards in terms of their aggressive properties. Lindzey and Goldberg (1953) and

TABLE 2

RANK ORDER OF NINE TAT CARDS FOR MANIFEST AGGRESSION

CARD	RANK									MEAN RANK
	1	2	3	4	5	6	7	8	9	
1						4	6	5	2	7.3
2		2		1	2	3	4	4	1	6.2
3BM			2	4	5	1	1	4		5.4
4	1	2	5	4	2	2	1			3.8
6BM		2	3	1	6	3	1		1	4.8
11	1	2	2	4		1	3	3	1	5.1
14				1	1	2		1	12	8.1
18BM	3	9	4	1						2.2
18GF	12		1	1	1	1	1			2.2

Stone (1956) found that cards 1 and 14 show little "aggressive pull." The latter author also found that cards 18BM and 18GF have especially strong "agressive pull" and that cards 3BM and 11 are of moderate "aggressive pull."

Further evidence that the cards are classified correctly is provided by data from the present study. The means on the aggression scale where the scores can vary from zero to six vary in the predicted direction. For the low aggressive cards the mean is .33, for the moderate aggressive cards it is 2.22, and for the high aggressive cards, 3.21.

Subjects

The subjects were forty volunteers from an introductory course in psychology and a senior course in psychology at UBC. Any subject who had told stories to TAT cards on any previous occasion was eliminated, as were those whose first language was not English. The age range of the subjects was from seventeen to forty-three, with a mean of 22.50 and a standard deviation of 5.09. Ten females and thirty males participated in the investigation.

At the time they were asked to volunteer, subjects were informed that they must be prepared to see the examiner for two sessions, each lasting about an hour. They were also told that the study involved an investigation of one of the major projective tests but were given no further information about the purpose of the study.

Procedure

All tests were given on an individual basis by the same examiner. Responses were recorded electrically and transcribed at a later time.

On the first administration all subjects were given conventional instructions, that is, they were instructed to give as dramatic a story as possible for each picture. (See Appendix II for complete instructions.)

At the second session, approximately one week later, every second subject (Group B, N=20) was asked to tell a different story to each card. The following paragraph was added to the reading of the original instructions: "You are urged to make no effort to recall your previous stories to the pictures. If one of the stories you told before comes to mind, simply put it aside and tell the next story that occurs to you." The other twenty subjects (Group A) were given exactly the same instructions as they had received on the first administration. If a subject asked if he should give the same or a different story, he was told, "That is up to you."

Scoring of the TAT protocols

The story protocols were scored on nine scales, and for each of the scales which was not completely objective an independent judge scored one story from each protocol in order to determine scoring agreement. The story selected from each protocol for this treatment was varied systematically, that is, card 1 for S-1, card

3BM for S-2, etc. This procedure was followed for the aggressive content scale, internal punishment scale, external punishment scale and similarity of plot. Scorer reliability coefficients are given in the text below and are always based on forty stories.

The records were scored for the following variables:

I. Aggression Effects

1. Stone's (1956) aggressive content scale. This is a weighted scale in which each aggressive response is categorized as involving a Death content, a Physical Aggression content, or a Verbal Aggression content. These content variables are weighted on a point system, as 3, 2 and 1 points respectively. Each response is also scored in terms of whether it shows active aggression or "potential" aggression. An action is scored as "potential" if the aggression is implied or placed in the future, or it may be a wish or idea that is not acted upon, for example, "He planned to kill her," or "He was thinking of suicide but changed his mind." If the action is "potential" only half the point credit is given. Scorer reliability for this scale yielded a Pearson product moment r of .92, $p<.01$.

2. Smith and Coleman's (1956) hostility control score. This score "...was obtained by dividing the number of hostile themes in a record which were not P (potential) scores by the total number of hostile themes produced" (p. 328). This score represents the degree to which the hostile feelings in the subject's hostile themes were acted out in the story as overt hostility.

3. Purcell's (1956) external punishment score. This score was arrived at by "...summing the frequency of such themes as the following when they were directed toward the hero: assault, injury, threat, quarreling, deprivation of some privilege, object or comfort, domination, physical handicap, such as blindness, etc., rejection." (p. 450). This scale reflects the subject's anticipation of extrapunitive aggression. Scorer reliability yielded a Pearson r of .70, $p<.01$.

4. Purcell's (1956) internally based punishment score. This score included "suicide, self-depreciation and feelings of guilt, shame or remorse" (p. 450). It is thought that this scale measures the subject's degree of anticipation of internal punishment. The Pearson r for scorer reliability on this scale was .94, $p<.01$.

II. Anxiety effects

A. Freezing Effects

1. Briefness, or the total number of words in the story, (Lindzey and Newburg, 1954). This score was merely a count of the number of words in each story, with the expectation that briefness will be associated with anxiety (Mandler et al., 1957).

2. Number of adjectives per 100 words, (Lindzey and Newburg, 1954). The total number of adjectives was divided by the total number of words, and a negative relationship with anxiety is postulated (Mandler et al., 1957).

B. Conflict Effects

1. Distress, a revision of the scale developed by Thomson (1960). One point was given for each fragment (where a sentence was left incomplete in meaning), and for each shift (where the subject started word or sentence and shifted before the utterance was finished). These points were added together for each story. A high score is indicative of anxiety.

2. Vagueness and hesitation, (Lindzey and Newburg, 1954). For each story a count was made of the number of statements showing either vagueness or hesitation, or both, for example, "I'm not sure," I don't know," "I can't tell."

III. Memory Effects: A global judgement was made on whether the two stories told by any subject to a single card were the same or different. This judgement was based on the plot of the story, rather than on any drive content. (See Appendix III for instructions given to the independent judge.) There was 95% agreement between the two judges for forty stories.

Statistical Analysis

Since the range of scores on most of the variables was very limited, tetrachoric correlations, calculated by the method described by Edwards (1954), were used in most cases to compute the internal consistency and repeat reliability coefficients. However, where the range was sufficiently great (Briefness, Number of adjectives per 100 words, and Hostility Control) product moment

correlations were calculated. Separate analyses were made for the cards in terms of the three levels of ambiguity and for the cards as a whole.

All p levels for the reported correlations are for two-tailed tests of significance. In the case of Pearson product moment r's with a sample size of 40, the obtained r must be .31 to be significant at a p of .05 and .40 to be significant at a p of .01. Comparable values with an N of 20 are .44 and .56, respectively. With respect to the tetrachoric correlation coefficients, their significance from zero was established by evaluating the significance of the corresponding chi squares. Thus, for a sample size of 40, the obtained tetrachoric correlation coefficient must be .60 to be significant at the .01 level and .47 at the .05 level. Similar values with an N of 20 are .79 and .64, respectively.

CHAPTER IV

RESULTS AND DISCUSSION

Repeat Reliability

It is evident from Table 3 that the repeat reliability coefficients for Group A (subjects given conventional instructions on both occasions) are fairly substantial in most instances. It will be noted from Table 3, however, that, since 81.7% of the stories were essentially the same, these relatively stable results are probably more a measure of memory effects than of reliability.

On the other hand it can be seen from Table 3 that the reliability coefficients for Group B are very low, with the exception of Briefness, Vagueness and Distress. The reliability of .64 for Briefness is not surprising in view of the established stability of word fluency. The correlations for Vagueness and Distress are probably spuriously high due to the fact that the scores on these scales were zero for the majority of subjects.

It is interesting to note that even when subjects were specifically asked to give <u>different</u> stories to the cards, over one-quarter of the stories were essentially the same. This would appear to cast some doubt on McClelland's hypothesis that "making a certain associative response tends to introduce resistance to give it again" (1958, p. 20).

Table 4 shows the repeat reliability coefficients obtained

by analysing the cards in terms of the three levels of ambiguity. The scores on the low aggressive cards for Session I were correlated with the low aggressive cards on Session II for each subject, the moderately aggressive with the moderately aggressive, and the high aggressive with the high aggressive on both sessions.

This procedure was designed to test Kagan's hypothesis (1955 that content categories reported to stimuli ambiguous for that content are less likely to be stable than those produced by stimuli which suggest that content. It is evident from Table 4 that the reliability of the three aggressive scales (aggressive content, external punishment and internal punishment) does not covary with the drive structure of TAT cards. This finding is not consistent with Kagan's hypothesis.

Internal Consistency

Internal consistency was evaluated by correlating the scores obtained on the first session for all subjects in terms of the level of ambiguity. The two low aggressive cards were correlated with the medium aggressive cards, the low aggressive with the high aggressive, and the medium aggressive with the high aggressive. These results are shown in Table 5. It is perhaps noteworthy that there is a slight trend for the medium aggressive cards and the high aggressive cards to correlate more highly than for the low aggressive cards to correlate with the medium aggressive cards or for the low aggressive cards to correlate with the high aggressive cards. In general, however, the correlations are somewhat low, with the exception again

TABLE 3

REPEAT RELIABILITY FOR SIX TAT CARDS
(N = 40)

SCALE	GROUP A (N = 20)	GROUP B (N = 20)
Aggressive Content	.59#	.00
External Punishment	.61##	.30
Internal Punishment	.81##	.02
Hostility Control	.17	-.41##
Briefness	.61##	.64##
Adjectives per 100 words	.19	.11
Distress	.28	.81##
Vagueness	.71##	.90##
Similarity of Plot	81.7% Same	25.8% Same

Note: Briefness, Adjectives per 100 words, and Hostility Control are product moment r's; all others are tetrachoric correlations. The same applies to the other tables.

\# $p < .05$
\## $p < .01$

TABLE 4

REPEAT RELIABILITY COEFFICIENTS WITH
CARDS OF VARYING AMBIGUITY TAKEN SEPARATELY

SCALE	Low vs. Low		Medium vs Medium		High vs. High	
	Group A	Group B	Group A	Group B	Group A	Group B
Aggressive Content	.07	.00	.95##	.30	.47	.30
External Punishment	.88##	.61	.62	.62	.94##	.16
Internal Punishment	.48	.00	.46	.27	.41	.20
Briefness	.66##	.49#	.55#	.60##	.70##	.77##
Adjectives per 100 words	.32	-.06	.23	.02	.17	.08
Distress	.80##	.32	.37	.65#	.61	.60
Vagueness	.63	.71#	.46	.62	.86##	.51
Similarity of Plot	85% Same	20% Same	82.2% Same	40% Same	77.5% Same	17.5% Same

Note: Hostility Control is not included in this table because this scale gives just one figure for each subject. This also applies to Table 5.

\# $p<.05$
\## $p<.01$

of Briefness and Vagueness. The size of these correlations indicates the necessity for caution in the use of additive treatment of scores from different TAT cards. Evidence must still be provided that similar responses to different cards are tapping similar psychological processes before a summation of scores from different cards can be justified.

TABLE 5

INTERNAL CONSISTENCY OF SIX TAT CARDS
IN TERMS OF LEVEL OF AMBIGUITY
(N = 40)

Aggressive Content	
Low vs. Medium	.16
Low vs. High	.21
Medium vs. High	.46
External Punishment	
Low vs. Medium	.04
Low vs. High	.04
Medium vs. High	.30
Internal Punishment	
Low vs. Medium	.48#
Low vs. High	.40
Medium vs. High	.04
Briefness	
Low vs. Medium	.48#
Low vs. High	.58##
Medium vs. High	.68##
Adjectives per 100 words	
Low vs. Medium	.03
Low vs. High	.08
Medium vs. High	.08
Distress	
Low vs. Medium	.18
Low vs. High	.68##
Medium vs. High	.03
Vagueness	
Low vs. Medium	.60#
Low vs. High	.32
Medium vs. High	.67##

# $p < .05$
$p < .01$

Comparison of Present Findings With Previous Research

It is difficult to compare the findings of the present study with previous work since the design of this study is quite different from most of the others. Only Lindzey and Herman (1955) instructed their subjects to tell different stories to the pictures on the second administration, thus minimizing memory effects which are apparently very strong. They did not score their stories for aggression in the repeat reliability part of their investigation, so no comparison can be made on this basis. However, their "abasement" can be compared reasonably with the internal punishment scale, and "dominance" and "rejection" with the external punishment scale used in the present study. In all cases the correlations they report are higher than those obtained in this study.

In measurements of internal consistency there is more agreement between the coefficients obtained by this investigator and those reported by others. Lindzey and Herman's (1955) reported coefficient of .29 for aggression is close to those obtained in this study (.16 for low aggressive content cards vs. medium, .21 for low vs. high and .46 for medium vs. high). These figures are also fairly close to the .34 for aggression reported by Child et al. (1956), and the .43 for aggression reported by Wylie et al. (1963).

Implications of Present and Past Findings

On the basis of the specificity hypothesis suggested by Kagan (1960) and accepted in this present study, the results presented

here could not be generalized beyond the two variables measured, that is, aggression and anxiety, as measured by the specific scales used. It seems safe to say, however, that the temporal and internal stability of these two variables in relation to the TAT is quite low.

In evaluating psychometric data on projective tests there seem to be two common approaches: one either recommends that the test be sent into oblivion, or else one points out the proven clinical value of the test, recommends caution in its use, and advocates further research to account for the lack of satisfactory levels of psychometric excellence. For this investigator the latter seems the more reasonable approach. The TAT has certainly established its clinical usefulness, but in view of the contradictory evidence on its stability, both internally and over time, well-designed research on these aspects of the test would be valuable.

With such low levels of internal and temporal stability caution is obviously required in research with the TAT. Rigorous controls are necessary since any observed changes, for example, pre- and post-therapy changes, may be attributed to the unreliability of the instrument. In clinical practice it seems that the usefulness of the test must continue to be based on the skill and experience of the practitioner.

Further profitable research could certainly be done on the McClelland-Atkinson hypothesized cyclical alternation. Is a third administration of the TAT more consistent with the first administra-

tion than is the second? Such research would have to pay special attention to the minimization of memory effects since these obviously loom large in repeated administrations of the test.

In using the instructions to tell a different story which were employed in the present study the question arises as to what these instructions mean to the subject. The impression of this investigator is that this meaning ranges from "Tell me another story if you can think of one," to "The last story you told was not satisfactory and I would certainly hope that you can do better than that." The motivational state of the subject will almost certainly vary, depending on his interpretation of the instructions.

CHAPTER V

SUMMARY

The purpose of the present study has been to evaluate the repeat reliability and internal consistency under short-term conditions of several indices of aggression and anxiety as measured by the TAT.

Three main questions were posed and investigated as follows:

1. What are the effects of varying the instructions on the second administration of the test? One group of subjects was given standard TAT instructions at both administrations, while a second group was asked to tell a different story to each card. This procedure was designed to control and study the influence of memory effects.

In the first group it was found that over 80% of the stories given were essentially the same on both administrations of the test, and even in the second group 25% of the stories were the same. It is thus apparent that memory effects are very strong and must be controlled in any repeat reliability investigations of the TAT. The relatively high repeat reliability coefficients for the first group are probably more of a measure of memory effects than of reliability per se, since the coefficients for the second group are very low.

2. What are the effects of varying levels of card ambiguity for a given drive (in this case, aggression) on the temporal stabi-

lity of that drive. Following Kagan's hypothesis (1955), it was predicted that temporal stability of fantasy variables would be a direct positive function of the drive structure of TAT cards. The findings of this study were not consistent with this hypothesis.

3. What is the degree of internal consistency of aggression and anxiety indices on the TAT? The internal stability was evaluated by correlating the scores obtained on the first session by all subjects in terms of the level of ambiguity. These correlations were also quite low, indicating the need for caution in using an additive treatment of scores from different TAT cards.

In all aspects of this study a specificity hypothesis, as suggested by Kagan (1960), has been assumed. This hypothesis asserts that no meaningful statements can be made about the TAT reliability or internal consistency in general, but only about specific variables. This hypothesis was accepted because of the variation in the few correlations reported in the literature on both internal consistency and repeat reliability of the TAT. Thus the results of this study could not reasonably be generalized beyond the two variables studied, namely aggression and anxiety.

REFERENCES

Atkinson, J. S., Studies in projective measurement of achievement motivation. Univ. of Michigan. Abstract in Univ. Microfilms, vol. X, no. 4; Publication no. 1945., 1950.

Auld, F., Eron, L. D., and Laffal, J. Application of Guttman's scaling method to the TAT. Educ. psychol. Measmt., 1955, 15, 422-435.

Child, I. L., Frank, K. F., and Storm. T. Self-ratings and TAT: their relations to each other and to childhood background. J. Pers., 1956, 25, 96-114.

Edwards, A. L. Statistical methods for the behavioral sciences. New York: Rinehart, 1954.

Haber, R. N., and Alpert, R. The role of situation and picture cues in projective measurement of the achievement motive. In J. W. Atkinson (Ed.) Motives in fantasy, action and society. Princeton, N. J.: Van Nostrand, 1958, pp. 644-663.

Jensen, A. R. Thematic apperception test. In O. K. Buros (Ed.) The fifth mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1959, pp. 310-313.

Kagan, J. The stability of TAT fantasy and stimulus ambiguity. J. consult. Psychol., 1959, 23, 266-271.

Kagan, J. Thematic apperceptive techniques with children. In Projective techniques with children. New York-London: Grune & Stratton, 1960, pp. 105-129.

Lesser, G. S. Custom-making projective tests for research. J. proj. Tech., 1961, 25, 21-31.

Lindzey, G. Thematic apperception test: the strategy of research. J. proj. Tech., 1958, 22, 174-180.

Lindzey, G., and Goldberg, M. J. Motivational differences between males and females as measured by the Thematic Apperception Test. J. Pers., 1953, 22, 101-117.

Lindzey, G., and Herman, P. S. Thematic apperception test: a note on reliability and situational validity. J. proj. Tech., 1955, 19, 36-42.

Lindzey, G., and Newburg, A. S. Thematic apperception test: a tentative appraisal of some "signs" of anxiety. J. consult. Psychol., 1954, 18, 389-395.

McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. The achievement motive. New York: Appleton-Century-Crofts, 1953.

McClelland, D. C. Methods of measuring human motivation. In J. W. Atkinson (Ed.) Motives in fantasy, action, and society. Princeton, N. J.; Van Nostrand, 1958, pp. 7-42.

Mandler, G., Lindzey, G., and Crouch, R. G. Thematic apperception test: indices of anxiety in relation to test anxiety. Educ. psychol. Measmt., 1957, 17, 466-474.

Purcell, K. The TAT and antisocial behavior. J. consult. Psychol., 1956, 20, 449-456.

Smith, J. R., and Coleman, J. C. The relationship between manifestations of hostility in projective tests and overt behavior. J. proj. Tech., 1956, 20, 326-334.

Stone, H. The TAT aggressive content scale. J. proj. Tech., 1956, 20, 445-452.

Thomson, M. C. The effects of stimulus deprivation on stories told to TAT descriptions. Unpublished MA thesis. U. Brit. Col., 1960.

Tomkins, S. S. The thematic apperception test. New York: Grune and Stratton, 1947.

Wylie, R. C., Sisson, B. D., and Taulbee, E. Intraindividual consistency in "creative" and "memory" stories written for TAT pictures. J. consult. Psychol., 1963, 27, 145-151.

APPENDIX I: Instructions to judges

GENERAL INSTRUCTIONS

The purpose of this study is to scale nine different pictures along a dimension of hostility (aggression). For the purposes of the scale you are asked to rank order the pictures in terms of the amount of hostility expressed in them. Base your judgement on the factor of hostility only.

You are to regard the pictures as social scenes illustrated in a story book.

Any of the following acts, thoughts or attitudes should be viewed as illustrating what we mean by hostility:

Physical hostility acts, such as killing, assaulting, combative, destructive, shooting, hitting, self-injury:

Hostile attitudes, such as being malicious, embittered, hating quarrelsome, domineering, irritable, scorning, grouchy, surly, resentful; and

Verbal hostility, such as being venomous, abusive, threatening, over-critical, argumentative, quarrelling, cursing, blaming, ridiculing and lying.

SPECIFIC DIRECTIONS FOR RANKING PICTURES

The pictures are spread out in front of you in a random fashion. You are to rank order them according to their degree of hostility. Examine all the pictures carefully before making any rankings.

Rank as number one the picture which expresses the most hostility, anger or aggression. Next, find the second most hostile picture and place it beside the first; then the third most hostile picture and place it by the second and so on down to the ninth picture that is least aggressive, which will appear at the extreme right.

After you have laid out the whole sequence before you, you may wish to change the order of the pictures. You may change the order of the pictures as many times as you like to obtain your final ranking.

Thank you for your cooperation.

APPENDIX II: Instructions to all S's on first TAT administration

"I am going to show you some pictures, one at a time, and I want you to make up as dramatic a story as you can for each. 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. Speak your thoughts as they come to your mind. You can make up any kind of story you please. Let yourself go freely. Do you understand? I want you to speak clearly so I can hear every word."

APPENDIX III: Instructions to independent scorer on evaluating similarity of plot

You are asked to judge whether the two stories told by any subject to a single card are essentially the same or different. The concern here is with the manifest plot content of the story, rather than any drive content. Exactness of language is not to be used as a criterion of similarity.

The following questions may be used in making your judgement:

1. Are the same characters involved in the story?

2. Is the hero the same?

3. Do the same things happen to the hero?

4. Is the outcome the same?

5. Are the thoughts and feelings attributed to the characters essentially the same?

The stories should not be considered different because of the addition or subtraction of details; however, if the first story is simply recalled, and then the subject proceeds to give the "next chapter" this would be a different story.