Faking Good on the Thematic Appreception Test

Roger S. Arnold
Loyola University Chicago

Follow this and additional works at: https://ecommons.luc.edu/luc_theses

Recommended Citation

This Thesis is brought to you for free and open access by the Theses and Dissertations at Loyola eCommons. It has been accepted for inclusion in Master's Theses by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License. Copyright © 1974 Roger S. Arnold
"FAKING GOOD" ON THE THEMATIC APPERCEPTION TEST

by

Roger S. Arnold

A Thesis Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Master of Arts

January

1974
ACKNOWLEDGMENTS

The author wishes to express his gratitude to Dr. Ann Heilman and Dr. Leroy Wauck whose guidance and criticism made it possible to complete the present study; to Dr. Donald Tyrell and his wife Barbara, to John, Patricia, and Mary Dickmann whose support and encouragement helped me over the difficult times; to Mrs. L. Evon Jarvis who typed the manuscript; and to the members of the Society of the Divine Word under whose auspices the time and resources were offered to make the study possible in the first place.
LIFE

Roger S. Arnold was born on June 28, 1933, in Pittsburgh, Pennsylvania.

He received a Bachelor of Arts degree in Philosophy from Divine Word College, Epworth, Iowa in June, 1957, a Bachelor of Theology from Divine Word Theologate, Techny, Illinois in June, 1961, and a Master of Arts degree in Canon Law from the Gregorian University, Rome, Italy in June, 1963.


He entered the graduate program in Experimental Psychology of Loyola University in January, 1970.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. REVIEW OF THE RELATED LITERATURE</td>
<td>9</td>
</tr>
<tr>
<td>III. METHOD</td>
<td>22</td>
</tr>
<tr>
<td>IV. RESULTS AND DISCUSSION</td>
<td>24</td>
</tr>
<tr>
<td>V. CONSIDERATIONS FOR FUTURE RESEARCH</td>
<td>37</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>40</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Summary of the statistical results of the Wilcoxon Matched-Pairs Signed Ranks test and the Mann-Whitney U test</td>
<td>25</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

The Thematic Apperception Test of TAT, was introduced by Christiana D. Morgan and Henry A. Murray (1935) as a method to explore unconscious thoughts and fantasies. They described a test that required subjects to tell stories associated with pictures, and they intended it as a device to speed up the progress of psychotherapy for patients who either could not afford or did not need complete psychoanalysis. Later Professor Murray (1938, 1943) expanded its scope by emphasizing that the TAT enables trained interpreters to reconstruct, on the basis of a subject's stories, his dominant drives, emotions, sentiments, complexes, and conflicts.

The test has had antecedents in the history of psychology. Having children react to pictures as a way of studying individual differences goes back at least as far as a Binet (Binet & Henri, 1896). Some of the older manuals in experimental psychology included exercises with ambiguous pictures to illustrate principles of apperception and sometimes stressed the significance of personal determinants (Sen, 1953). Early in the century, Brittain (1907) and Libby (1908) employed the storytelling technique for studying imagination in adolescents. These pioneer studies, which all dealt with such general facets of mental life as intelligence, apperception, and imagination, were not concerned with the picture-story method as an index to the individual personality.

A more direct precursor was the set of pictures L. A. Schwartz (1932) used as an adjunct in the psychiatric interviewing of delinquents.
Van Lennep devised his Four Picture Test in 1930 but did not publish for eighteen years (1948).

The principal credit, however, for the innovation still belongs to H. A. Murray, who in the 1930s was the moving spirit in developing new projective techniques at the Harvard Psychological Clinic for the systematic appraisal of college students. Of these techniques, most of which have fallen into disuse, initially the most promising and eventually the most far-reaching in influence was the TAT.

Although the TAT was at first slow in gaining wide acceptance, it is now a test that approximates the Rorschach in popularity and in the amount of research it has stimulated. At first it gained popularity only among clinical psychologists, but gradually it became a research tool in developmental psychology, social psychology, personality, and cross-cultural studies in anthropology. It is also used for personality assessment in the fields of counseling and industrial psychology. Paralleling the widespread uses of the test, the TAT has given rise to vast literature describing its many practical applications and theoretical underpinnings (Harrison, 1965; Murstein, 1963; Symonds, 1949; Tomkins, 1947).

The principal assumption underlying traditional TAT interpretation is that fantasies, experiences, desires, inner tendencies, attitudes, and conflicts have been projected onto the stimulus cards and the characters or objects depicted on them. However, this assumption has been seriously questioned by a number of TAT theorists. It has been noted (Tomkins, 1949), for example, that important fantasies and experiences are actually withheld
from TAT stories.

Further, it has been assumed by TAT users that the wishes, needs, personality styles, and motivation components revealed in the TAT stories constitute material that cannot be secured without the test. This assumption holds that respondents are either unwilling or unable to talk about them in the ordinary psychodiagnostic or therapy setting. Whereas Murray (1943) only hoped that the subject would become so involved in the task that he forgets his sensitive self and the necessity of defending it against the probings of the examiner, most TAT investigators have readily assumed that these unconscious components of personality cannot be obtained in another fashion.

Finally, two important assumptions are made in interpreting the test. It is assumed by some that test scores are related to overt behavior; and that it is sensitive to varying motivational states.

The task, then, confronting the interpreter of TAT protocols is to relate the small segment of behavior reflected in a subject's story to some facet of his personality or to some behavioral outcome. To provide these relationships, at least ten major interpretive schemes are available, and this does not include the multitude of home brewed interpretive systems developed by clinicians in their daily practice.

The last mentioned problem provides the jumping off point for the present study.

In a recent book, Fiske (1971) makes the following statement:

"To advance the science of personology, intensive effort must be devoted to each major construct, to delineating it explicitly and
systematically, and to creating measuring procedures conforming to the blueprints derived from such conceptualization (p.14)."

Later he says:

"The most serious matter is the extent to which the indices obtained in a testing situation serve our purpose by capturing the essence of the construct being measured. ...This is the matter of validity, the crucial aspect of all measurement, the objective toward which all other concerns are directed. The other central concept in psychometric theory, reliability, deals with means to that end, with the degree to which one can generalize from the indices obtained under specified procedures to indices that might be obtained under procedures differing in one or more respects (p. 128)."

Is the TAT the kind of measuring procedure that fulfills the desire expressed by Fiske (1971)? Does it measure a major personality construct? Is it valid and reliable?

Arnold (1962) published a work on the TAT which deals concretely with the above questions. She presented a method of analyzing stories for assessing the motivation of normal people and found that it was useful for discovering positive and negative motivation both through the scoring system she devised and through clinical evaluation. She used the already existing test developed by Morgan and Murray (1935), but proposed both theoretical and practical changes to make it more effective. According to Murray, a man is likely to reveal his motivation, that is, his needs, wishes, hopes and fears, while interpreting an ambiguous social situation, such as is portrayed in each of the TAT pictures. In the first version, the instructions were simply to interpret each picture, to guess what went before and what was the final
outcome. Increased experience gradually modified the instructions. Now the subject is asked to tell a story about each picture.

Arnold criticized the assumptions upon which the test had been based. The theory of personality used by TAT interpreters had usually been psychoanalytic theory. This proposed that instinctive drives or impulses are the real motivating forces of man which are modified by ego-processes in overt behavior, but reveal themselves in fantasy. She pointed out that drives and affects that are assumed to be projected upon the story characters are not a sure guide to the kind of motivation that leads to action in everyday life. Her suggestion was that simply taking the "themes" in the TAT indicates preoccupation but not genuine striving. Actually, the storyteller's motivation is illustrated not by the story themes but by the story outcome and the way the story is told.

This theoretical difference led her to criticize the existing scoring systems. The underlying assumption of the scoring methods was the notion that the story is an aggregate of themes, and these themes must be isolated before scoring is possible. Themes may be categorized in a variety of ways, from Murray's (1943) needs and press, Tomkin's (1947) vector and level analysis, to Kagan's (1961) affect states; but in every case, the scoring is based on isolated parts of the story.

There are several interpretive techniques using the outcome in addition to story themes (McClelland et al., 1953; Eron, 1951; Hartman, 1951). However, the outcome is not just another theme like aggression, hostility, affiliation and the like. The outcome caps the plot, and the plot integrates
various themes into a unified whole. A story is a creative reorganization of past sense impressions, a new product of human imagination. The story has a meaning which cannot be discovered from the meaning of the individual themes into which it can be analyzed. Whatever score we may assign to such parts, and however we may manipulate or categorize the elements into which we have divided the story, we are disregarding the story so long as we deal with themes rather than with plot and outcome. If it were possible to score what the story is saying, and include all there is to the story, we would discover a new dimension in the TAT which might allow very different correlations with behavior.

To solve that problem Arnold (1962) proposed that each story be condensed into an import that leaves out incidental details but preserves the kernel -- the meat of the story. When all the imports are read in sequence, a picture of the individual emerges that does portray his attitudes and intentions for action. Every story makes a point, expresses a conviction. It describes an action that may be headed for success or failure, may exemplify cooperation or hostility, may be an attempt to cope with adversity or betray spineless acceptance of whatever may come. In every case, emotions may influence the action, but the outcome is primarily an expression of the storyteller's convictions, gathered from experience and reflection. The plot sets a problem, the outcome solves it. Both the type of problem a man sets for himself and the kind of outcome he prefers are characteristic for him. Every story has a moral though that moral may not indicate very high ethical principles. It is possible to score such story imports objectively and arrive at a final score which will indicate a man's positive or negative motivation.
Arnold's (1962) analysis and criticism of existing methods of TAT interpretation offers a solution to the long-standing ambiguity about the scientific usefulness of data from TAT testing.

Sargent (1945) says that the variety and richness of material which the projective methods provide is at once the delight of the clinician and the despair of the experimentalist. The research worker who attempts to use any of these methods is immediately impressed both with their infinite possibilities for interpretation and insight, and seemingly insurmountable difficulties in the way of scientific treatment of the data. The problems of quantification and standardization have been the chief source of controversy, not only between advocates and opponents of the methods but among the enthusiasts themselves. The application of standard methods for estimating reliability and validity further complicates the problem of quantification. In psychometrics, the usual checks on reliability have been the split-half technique, and correlations either between repetitions of the same test or between alternate forms. None of these methods, according to her, are wholly satisfactory for projective tests.

In Arnold's (1962) system, however, each story is crystallized into a single item which can be objectively scored. These items can be readily compared either on the basis of a split-half technique or test-retest. Thus, the objection loses much of its force.

Tomkin's (1942) repeating the TAT daily with a group of subjects found that 20 sessions were required to bring out all the significant themes for one person. This finding seems to be enough to indicate that high reliability can hardly be expected upon just one repetition of the test.
Again, this objection holds true only if the test is analyzed in terms of themes. If it is analyzed in terms of imports the results of a re-test will be reliable.

In an overview article, Harrison (1965) has the following:

"The experiences with the TAT are so diversified and the literature so extensive and variegated that it is possible to find precisely opposite opinions being expressed on a broad range of topics. The TAT is said to be valid and not valid, reliable and not reliable, predictive of overt behavior and not predictive of overt behavior, useful in research and not useful, easily faked and not easily faked, good for diagnosis and not good for diagnosis, economical and not economical, capable of plumbing the unconscious depths and shallower than other projectives. There is also radical disagreement between the subjective, or qualitative approach on the one hand and the objective, or quantitative, approach on the other. Furthermore, it is possible for some writers to prove to their own satisfaction either side of an antinomic proposition by presenting what amounts to a lawyer's brief through either insufficient exposure to the relevant literature or careful selection of evidence (pp. 563-564)."

Each one of the polarities listed above would be valuable and interesting for study using Arnold's (1962) system for TAT analysis. The main questions to be asked in this study, however, are these: a) does Arnold's method of TAT interpretation and scoring yield reliable results in a test-retest situation; b) is the TAT reliable in a "fake good" situation?
Anastasi (1961) says that the reliability of a test refers to the consistency of scores obtained by the same individuals on different occasions or with different sets of equivalent items. The concept of test reliability itself has been used to cover not one but several aspects of score consistency.

The most obvious method of finding the reliability of a test is by means of a retest, or repetition of the identical test on a second occasion. The reliability coefficient in this case is simply the correlation between the scores obtained by the same subjects on the two administrations of the test.

Although apparently simple and straightforward, this technique presents difficulties when applied to most psychological tests. Practice will probably produce varying amounts of improvement in the retest scores of different individuals. If the interval between retests is fairly short, the subjects may recall many of their former responses. Thus, the scores on the two administrations of the test are not independently obtained and the correlation between them will be spuriously high. Only tests that are not appreciably affected by repetition lend themselves to the retest technique.

One way of avoiding the difficulties encountered in retest reliability is through the use of equivalent forms of the test. The subjects can then be tested with one form on the first occasion and with another, comparable form, on the second. The correlation between the scores obtained on the two forms represent the reliability coefficient on the test.

Split-half reliability. From a single administration of one form of a test it is possible to arrive at a measure of test reliability by various split-half procedures. In such a way, two scores are obtained for each
individual by dividing the test into comparable halves. Split-half re-
liability provides a measure of equivalence, or adequacy of item sampling.

Reliability of the TAT

Ross (1965) says that the conventional reliability methods that have
been evolved for psychometric tests, such as measures of ability and
achievement, cannot be applied in unmodified form to projective techniques
like the TAT which are based on entirely different principles. Thus split-
half reliability is not suited because, unlike psychometric items, the pictures
were designed to yield psychologically different, not equivalent data.
Arnold's solution to this problem will be given below.

Regarding the retest method of gauging temporal stability, clinical
care is not with whether the individual over a period of time relates
the same narratives but with the significant material that emerges from
story analysis. Giving the same pictures to subjects after a short interval
creates difficulties in motivation; also, the pictures are not phenomenally
the same. McClelland (1958) called attention to the demonstration years ago
by Telford (1931) that making an associative response tends to induce re-
sistance to making the same response immediately a second time. Tomkins
(1947) had something like this in mind when he pointed out that if a person
does not appreciate a joke as much when told a second time, this does not
indicate that the original response was meaningless or was a poor basis for
making inference about the laughter. So if a subject is shown the pictures
a day later with instructions to relate the same stories, story consistency
will probably be high but will be a test of memory only; if the instructions
are to give different stories, story consistency will be much less; and if the
instructions are left open, the results will depend on the subject's
interpretation of what is expected of him. In neither event is anything meaningful demonstrated about the reliability of the instrument. Moreover, if retesting is done after a considerable interval, the psychological state of the subject is likely to have changed, and what comes out in the form of stories should reflect these changes without being considered unreliable. Moods and feelings, problems and preoccupations, it must be apparent, are subject to more fluctuation than knowledge and abilities. The degree to which these methods are irrelevant will depend on whether the researcher or clinician is employing the TAT as a psychometric instrument or for global analysis. If he scores stories and tries to use the TAT as a psychometric test, then to some extent he must be judged by psychometric standards, even though the split-half and retest methods are not as meaningful as with statistically constructed tests.

The above qualifications on reliability studies of the TAT are proposed by Ross (1965). They are less pertinent to studies done using Arnold's method since the imports derived from the stories are, by and large, psychometrically equivalent. However, the results of the study must always be viewed in the light of the difficulties he mentions.

**Split-half reliability**

Possibly because of the common recognition that internal consistency as usually calculated is not appropriate for the TAT, there have been relatively few split-half coefficients reported (Bialick, 1951; Calogeras, 1958; Child et. al., 1956; Lesser, 1957; Lindzey & Herman, 1955; Sanford, 1953). With the exception of the Child study, which had a range from −.10 to .44, the resulting coefficients fell between .40 and .85. Most of the
Results were obtained from varying numbers of the Murray cards with cards and press as the usual measure. More pertinent for the problem internal consistency was the demonstration by Palmer (1952) that persons can be identified from their stories regardless of picture.

Jensen (1959), however, says that the widely held misconception that split-half or internal consistency reliability is meaningless in the TAT. He criticizes the Child (1956) and Lindzev (1955) studies by saying 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. It would be like adding together pounds, gallons, and inches. Each card seems to be a unique test in itself and is correlated little, if at all, with other cards. This fact casts serious doubt on the validity of many methods of TAT interpretation.

Test reliability

Five days a week, over a period of ten months, Tomkins (1942) administered a new picture to a single subject. He also gave the entire test three times at three-month intervals and once when the subject was under the influence of alcohol. He found that the main themes all appeared in the first full administration of the test and that in spite of attempts to make the stories different, they reoccurred on subsequent administrations. Other reports, where observations on temporal stability were incidental, have indicated consistency over time but do not give statistical data (L. Fisher, 1958; Greenwald, 1959; Symonds & Jenson, 1961).
Lindzey & Herman (1955) cited Tomkinc's (1942) study but complained that he did not give sufficient details to make it possible to evaluate his results. They also note that the area of repeat reliability, i.e. consistency of measurement over a period of time, has been little explored. They give a number of reasons for this. Stability of measurement over time is a somewhat mixed blessing as many investigators assume that, given any appreciable period of time, changes take place in the individual that should influence test response. Thus, although the test is customarily used to get at enduring characteristics, it also is assumed to vary with temporary states or motives. Further, it is clear that the time and effort involved in collecting data for repeat reliability studies is not commensurate with the lowliness of this empirical goal among most clinicians. Finally, there are procedural problems posed by the contribution of memory to the stories told on the second administration of the test that make this an awkward empirical problem.

They found, however, that considering the probable variation introduced by changes taking place in the subjects during the two-month interval, in addition to the variance contributed by imperfect rater-reliability, the consistency observed is relatively satisfactory. The higher correlations compared quite favorably with many inter-rater reliabilities reported for this test.

In other studies where statistics are available, moderate to high correlations are reported (H. H. Morgan, 1953; Tomkin's, 1947; Weisskopf & Dunlevy, 1952), but sometimes the consistency is poor (M. O. Bradley, 1957; Krumholtz & Farquar, 1957). The time intervals have usually been weeks,
sometimes months, whereas in psychometric studies retesting is usually done a few days apart.

Reliability of the Motivation Index of Story Sequence Analysis

As was said above, Arnold (1962) devised a system called Story Sequence Analysis for measuring motivation. Her motivational theory is explained primarily in Emotion and Personality (Vol. II, 1960) and her method of using the TAT is set forth in Story Sequence Analysis (1962).

(From now on Arnold's system will be referred to as the SSA of the TAT, or simply SSA.) The unit of interpretation in the SSA is the story, rather than the themes of the stories. The critical process in the method is extracting the import of each story. The import is a condensation of a story which preserves the kernel of the story and leaves out incidentals. Each story makes a point, and when this point is extracted in the import, and the imports are strung together in a sequence, a picture emerges of the person's motives and intentions to actions.

"If each story is an imaginative exploration of various problems and their possible solutions, we must try to isolate what it is the storyteller is trying to say. What he says about the picture will reveal his convictions: what could be called the "moral" of the story. When this moral is applied to the storyteller's subjective circumstances, we arrive at the import (the meaning or significance) of the story. Once the import of each story is set down in sequence, it becomes possible to follow the storyteller's trend of thought, which reveals his habitual dispositions, the way he evaluates human actions, and the circumstances of man's life. The story import will show how the storyteller thinks people usually act and how he
feels they should act; what actions he thinks right and wrong; what will lead to success, in his opinion, and what to failure; what can be done when danger threatens, and what things to strive for. In short, the story imports, taken in sequence, give a connected statement of the storyteller's principles of action, his motivation pattern (Arnold, 1962, p. 51)."

Skill in taking imports from stories is learned, for the import is objective in that it is abstracted as well as possible without adding from subjective interpretations. However, because of the examiner's focus on a particular problem around which the story is spun, imports as well as the sequence in which they are embedded may vary somewhat from examiner to examiner. Some stories are straightforward and clear as to what the storyteller is trying to say. With others, the point of the story is less well made, outcomes are implied rather than stated clearly, stories border on monologues, and the import is difficult to score. Here skill in interpreting has to be developed, and when the stories are written with definite plots and outcomes, the meaning of the import is consistent when taken by different examiners although the particular wording of it may vary considerably. When the import is scored using the categories that have been devised, the variations in wording of the import become unimportant and the import is preserved in a unified score.

Scoring

There are two aspects of scoring: classifying the content of the story and giving it quantitative value. Imports fall into four broad categories:

I. Achievement, success, happiness, active effort and their opposites;
II. Right and Wrong; III. Human Relations; and IV. Reaction to Adversity. These categories were derived empirically from the records of high and low achievers, from effective and ineffective teachers, from offenders and non-offenders among Navy recruits, and efficient and inefficient executives in a government project. Over a period of ten years these categories of imports were organized into a scoring system with various headings and sub-headings to include the types of imports found in the stories.

Quantitatively, each story has four possible scores ranging from -2 to +2. These scores derive from the types of imports found in positively motivated persons as contrasted with those found in negatively motivated persons and were derived empirically from the contrasting groups mentioned above. The algebraic sum of all imports is transformed into a score based on the number of cards used and is called the Motivation Index. The M.I. is calculated according to the following formula:

\[
\frac{\pm n^o \times 200}{n^p} + 100
\]

Where \( n^o \) is the number of units obtained, the algebraic sum of the scores of the imports, and \( n^p \) is the number of units obtainable or possible. The \( n^p \) is four times the number of cards used: e.g. for a 20-story sequence, the \( n^p \) is 80. Arnold provides a table to make this transformation easy (Arnold, 1962, pp. 146-147).

Since every story is scored by reference to the same scoring criteria, and since the import derived from the story is scored and not the picture.
about which the story is told, every "item" in the test is equal to every other "item". This has been shown to be the case by H. L. Fagot (1961) who found that the two positive and two negative scores were evenly distributed among all the cards used in a sample of 252 records of twelve stories each. For this investigation, Fagot used the records of 99 teachers (Burkard, 1958), 100 college students (Garvin, 1960) and 53 seventh-grade children (Arnold, 1961).

Since the "items" measured by the intensity scale are equivalent, even one story should theoretically give a correct picture of the storyteller's intensity and direction of motivation if he were completely consistent and the scoring were completely reliable. Unfortunately, not many individuals are so completely consistent in positive attitudes that they would always tell stories with highly positive imports. On the other hand, it is fortunate that there are still fewer people who have such strong (and strongly entrenched) negative attitudes that they would reveal them in every story they tell. Each story, then, is considered as a sample of the storyteller's motivating attitudes, his habitual motivation. The intensity score is a measurement of this motivation. Demanding several stories is equivalent to applying repeated measurements. Instead of measuring various dimensions of the quality we want to measure, as is done in intelligence tests and most personality tests, many measures of one and the same dimension are taken, which is the storyteller's motivation, positive or negative. Highly consistent people will reveal essentially the same type of motivation in every sample and show high "reliability";
as do, for instance, the good and poor teachers. But in a group that represents the whole range of consistency, one has to take quite a few samples before he can be sure that he can strike an average that will fairly represent the individual's degree of consistency. Arnold (1961) found high correlations with the outside criterion when using 11-, 12-, 13-, and 20-story sequences.

Faking

Harrison (1965) says that one of the special merits attributed to the TAT is its relative immunity to deliberate misrepresentation, which is the bane of personality inventories. On the TAT a sophisticated person who is fully aware of a particular problem or conflict may keep it out of his stories by conscious intent but will often reveal his defensive handling of the conflict, and many times during the course of a full-scale test he will be unable to sustain his defenses sufficiently to keep the problem from finding at least indirect expression.

Weisskopf and Dieppa (1951) showed that hospitalized veterans could influence their stories in a favorable or unfavorable direction, so that the diagnostic conclusion of TAT interpreters were affected on some but not on all personality dimensions. Only three pictures were used, however, so one is left to speculate what the result would have been if more pictures had been used.

Dunlevy (1953) obtained the same results when examiners were unaware of the purpose of the experiment. When they were forewarned that some subjects were trying to make more, and others less favorable impressions, they were able to detect stories told under the different instructions with
some success. These conditions are closer to those in practical testing, where test behavior and circumstances surrounding the test, as well as biographical data, all combine to alert the examiner to possibilities of misrepresentation.

Davids and Pildner (1958) administered inventories and self-rating scales as well as projectives to two groups of college students. One group was taking tests as part of an actual job application, the other as participants in a research project. The job applicants obtained significantly higher adjustment scores than the research subjects on self-reporting devices but not on the TAT and other projective tests.

Ni (1959) was able to detect attempts at concealment for six college students from blind analysis.

Dollin and Reznikoff (1961) administered the Edwards (1957) Social Desirability Scales (short form) and six TAT cards to 48 psychiatric patients and 56 nursing students.

While patients with different Social Desirability tendencies showed significant parallel differences in their TAT stories, nursing students did not. Edwards maintains that social desirability values seem relatively constant regardless of age, sex, socio-economic level, and patient versus college student status. There was, thus, a population difference not predictable from the theory. There was no confirmation of an S.D. bias on TAT responses. Subjects with high scores on the S.D. scale did not consistently show high S.D. tendency in their stories.

In another study, Reznikoff (1961) evaluated the role of social desirability in the choice of TAT themes. He utilized the 60 TAT themes found to occur most often by Eron (1950) in his normative investigation.
The correlations between S.D. did not even approach statistical significance. These data suggest that S.D. does not contribute to TAT theme variance.

There have been no studies on faking using the Motivation Index of Story Sequence Analysis as a basis of measurement.

**Purpose of the present study**

The main purpose of this study, therefore, is to examine the consistency of the M.I. scores under conditions of "fake good" and compare them to the scores of a control group under normal testing conditions. It is assumed that SSA will be particularly resistant to faking good since Arnold found a phenomenon called the "Pollyanna" story to be characteristic of the low-achiever (i.e. poorly motivated individual). An unsophisticated subject might assume that a "happy ending" story would indicate a "healthier" personality and that a gory and aggressive one with a violent outcome would indicate the existence of psychological problems. However, easy, unfounded optimism gets a minus score in the SSA and in some cases negative endings get a plus score. So it is possible that a subject might actually lower her score by attempting to fake good.

The secondary purpose is to test the reliability of the M.I. under "test-retest" conditions. This will be done for both the normal and the "fake-good" group.

**Hypotheses**

1. Subjects taking the TAT under "fake good" conditions will have M.I. scores consistent with those they received under normal conditions.
2. The subjects' retest scores will correlate highly with those received on the first testing. This will hold for both conditions.

3. The "fake good" group will not differ significantly from a control group taking the TAT under normal conditions with regard to the M.I.
CHAPTER III

METHOD

Subjects

The subjects of this study were 61 unmarried college males residing in a preparatory seminary college. They were from 21 to 24 years of age. Thirty-seven subjects were used for the control test-retest group and twenty-four for the "fake good" group.

Procedure

The TAT was administered according to the instructions in Arnold's (1962) Story Sequence Analysis using thirteen cards (1, 2, 3BM, 4MF, 6BM, 7BM, 10, 11, 13MF, 14, 16, and 20 shown in that order). The test was given in a group setting using slides of the pictures projected onto a screen.

Before the first picture was shown these instructions were given:

"This is a test of your creative imagination. You will see thirteen pictures, one after another. As you look at each picture, write as dramatic a story as you can about it. Tell what has led up to the scene shown in the picture and what is happening now. What are the thoughts and feelings of the people in the picture? What will be the outcome?

Since we are interested in your creative imagination, be sure to tell a story with a plot and an outcome. Do not just describe the picture. Try to write a story and not a piece of conversation.

You will have seven minutes for each story. Be sure to write a story with a plot and an outcome. Do not just describe the picture. Try to write a story and not just a piece of conversation.

You will have seven minutes for each story. Be sure to write something about each picture. If you can't think of anything for one of the
pictures, write that down too. (Arnold, 1962, p. 49)"

After the first story has been written, the instructions were repeated stressing that a story with a plot and an outcome be written, giving what led up to the scene, what is happening in the picture, and how it will turn out.

Both groups were tested a month later. The control group was given the same instructions as at the first testing. The experimental group received the following instructions:

"You are to write the stories in such a way as to indicate that you are functioning much better psychologically than when you took the first test. For example, you can write as if you had received a very positive psychotherapeutic experience and this changed you for the better."

Measures

The measures were the Motivation Index scores derived from the imports of the TAT stories as proposed by Arnold's (1962) Story Sequence Analysis

The TAT records were scored by blind analysis, that is, without the scorer knowing to which group the record being scored belonged. The imports were taken and scored independently by two examiners. After the independent scoring was completed, the scoring was compared and it was found that there was 78% agreement between the two examiners. The discrepancies were usually due to oversights on the part of one the examiners with regard to the wording of the import. The differences were discussed until the examiners resolved the discrepancies.
Statistical Analysis

Since the rating scale of the Story Sequence Analysis scoring system has no equal intervals, it cannot be treated with parametric statistical procedures. The intensity of the scored import is indicated by -2, -1, or +1, +2 (Arnold, 1962). These levels of positive or negative motivation cannot be considered a metric scale, for when a rater changes from positive to negative scoring, it is not a step equal to changing from 1 to 2 on the same side of the scale. This intrinsic inequality of steps in the rating scale limits the type of statistics that can be used. Therefore, for this analysis non-parametric methods that depend on rank ordering are employed.

To test the correlations the non-parametric statistic chosen is the Spearman rank correlation coefficient, \( \rho \). It is a measure of association which requires that both variables be measured in at least ordinal scale so that the individuals under study may be ranked in two ordered series.

The Wilcoxon Matched-Pair Signed Ranks test is used to compare the pre- and post-test of each group because the scores of the subjects can be matched.

The Mann-Whitney U Test will be employed to compare the control group with the experimental group for in this case matching is not possible.

The structure and use of these statistical procedures is taken from Siegel (1956).
CHAPTER IV
RESULTS

Descriptive statistics

The median Motivation Index score for the control group on the pretest was 77 with a range of 113; on the posttest the median score was 81 with a range of 136.

The median M.I. score for the experimental group on the pretest was 62 with a range of 104; on the posttest the median score was 69 with a range of 82.

Correlational statistics

The correlation between the test and retest of both the control and experimental groups was found by using the nonparametric Spearman Rank Correlation Coefficient $r_s$. The results were the following: a) the correlation between the pretest and posttest of the control group was .58 which is significant at the .01 level; b) the correlation between the pretest and posttest of the experimental group was .46 which is significant at the .05 level.

Inferential statistics

The results of the Wilcoxon Matched-Pair Signed Ranks test and the Mann-Whitney U test are summarized in Table 1.
Table 1

A. Wilcoxon Matched-Pairs Signed Ranks Test

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental group:</td>
<td>T = 110.5 with</td>
<td>NS: p &gt; .05*</td>
</tr>
<tr>
<td>pre-post comparison</td>
<td>N of 22</td>
<td></td>
</tr>
<tr>
<td>Control group:</td>
<td>z = -0.322</td>
<td>NS: p &gt; .3745</td>
</tr>
<tr>
<td>pre-post comparison</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. Mann-Whitney U Test

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest: experimental</td>
<td>z = 1.434</td>
<td>NS: p &gt; .15</td>
</tr>
<tr>
<td>and control comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest: experimental</td>
<td>z = 0.332</td>
<td>NS: p &gt; .3707</td>
</tr>
<tr>
<td>and control comparison</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* For N of less than 25 the probability of T is given in Table G in Siegel (1956).
The Wilcoxon Matched-Pair Signed Ranks test is applied to comparisons between pre- and posttesting of each group because the scores of the subjects can be matched. The Mann-Whitney U test is used to compare the control group to experimental group before and after the treatment since matching is not possible.

The Mann Whitney U test showed that the difference between the control group and experimental group prior to the treatment was non-significant (p > .15). When it was applied to control and experimental group after the treatment the difference was again nonsignificant (p > .3707).

The Wilcoxon Matched-Pairs Signed Ranks test indicated that the difference between the pre- and posttest scores of the control group was nonsignificant (p > .3745). It also showed that there was no significant difference between the pre- and posttest scores of the experimental group (p > .05).

The hypotheses presented for examination in this study were all supported by the data.

1. Subjects taking the TAT under "fake good" conditions will have M.I. scores consistent with those they received under normal conditions.

If the subjects in the experimental group had successfully "faked good" on the second testing one would have expected the second set of score to be significantly higher than the first set. However, the Wilcoxon Matched-Pairs Signed Ranks test showed that there was no significant difference between the pre- and posttest scores (cf. Table 1) of the experimental group. Therefore, the data show that the scores under the "fake good" conditions are consistent with those received under normal conditions.
2. The subject's retest score will correlate highly with those received on the first testing. This will hold for both conditions. If the TAT is a reliable test one would expect the correlation between the test and retest to be high. This was the case for both the control and the experimental group. The correlation between the two testings of the control group was .58 which is significant at the .01 level; the correlation between the two testings of the experimental group was .46 which is significant at the .05 level.

3. The "fake good" group will not differ significantly from a control group taking the TAT under normal conditions with regard to the M.I.

It was first necessary to find out whether the control group differed from the experimental group under normal testing conditions. The Mann-Whitney U test was used to compare the pretest scores of the two groups and it was found that they did not differ significantly (cf. Table 1). On this basis it is reasonable to assume that the groups were roughly equivalent prior to the experimental treatment. Therefore, if the experimental group had been able to "fake good" successfully their scores should have been significantly better than those of the control group on the second testing. However, the Mann-Whitney U test showed that there was no significant difference between the two groups on the second testing. Thus the data support the hypothesis that the attempt to "fake good" did not appreciably affect the scores.

The over-all results of the study, then, supported two conclusions.
1. The TAT, interpreted and scored according to the SSA method is generally immune to deliberate attempts to "fake good."

2. The TAT is reliable in a test-retest situation.

There were three protocols in which the changes were greater than could be expected from the statistical analysis. One subject improved his score from 65 to 135 on the "fake good" and two subjects lowered their scores on the "fake good" from 112 to 77 and from 112 to 81 respectively. The latter differences can be explained by the "Pollyanna" story phenomenon discussed earlier. The subjects told a number of these stories on the "fake good" retest and consequently lowered their scores considerably. The drastic improvement is more difficult to explain. One possibility is that the subject was lucky enough or sophisticated enough to improve the stories without succumbing to the "Pollyanna" trap. Another possibility is that the student was influenced by a "contaminating" factor such as psychotherapeutic experience which improved his motivation. This is unlikely since the students at the college were not exposed to any psychotherapeutic interventions either individual or group during the month between the testings. They all reside at the college, by and large attend the same classes, and do not leave the campus except for short visits to a nearby town. So there is little chance of some outside influence which would have caused that large a gain. So it seems that in isolated instances the score of the SSA can be significantly improved under "fake good" conditions. This represents, however, a random deviation from what normally occurs. Generally speaking, on the retest the lower scores improved somewhat and the higher scores decreased.
This is consistent with the universal phenomenon of regression toward the mean.

2. The TAT is relatively reliable in a test-retest situation. Although the correlations for both the control and experimental groups were statistically significant the "fake good" group was at the .05 level and the control group was at the .01 level. This is accounted for by the discrepancies noted above. It seems that the lower correlations are particularly affected by the lowering of the M. I.'s due to the "Pollyanna" stories.

Relationship of the Results to Previous Studies

It is somewhat difficult to make exact comparisons between this study and others of its kind. The previous studies practically all employ Murray's (1943) method of interpreting and scoring the TAT. The SSA method is different both theoretically and practically from the "thematic" approach which is most in vogue. So the investigator will try to show how some of the difficulties encountered in previous studies could have been avoided or at least alleviated by SSA method.

Tomkins (1942) gave the entire test three times at three month intervals and once when the subject was under the influence of alcohol. He found that the main themes all appeared on the first full administration of the test and reoccurred on subsequent administrations. A direct comparison to these results is not possible since Tomkins bases his findings on the similarity of themes rather than on similarly scored imports. The main difficulty with Tomkin's approach is theoretical. A person could give similar themes over a period of time but could organize these themes into
stories with outcomes which would be motivationally quite different. Conversely, stories with different themes could get the same M.I. score and be motivationally equivalent. Two examples will illustrate this point. The first is from the protocol of a 21 year old female college student before and after psychotherapy; the second is from one of the students in the present study. All four stories are told to card 3BM.

"The woman is in jail and a guard opened the door, she thought he was going to let her out, but he came in and started slapping her around so she grabbed his gun and shot him." Import: When a person who you hope will free you starts to abuse you, you take revenge. Score: minus two.

"The woman in this picture is in a room in a jail, she can't even remember why she's there but she decided she wanted to leave. There are men in uniform who stand at the exits and she figures that they are guards. She had been out of her room and when she came back she found a gun lying on the floor in the corner. She wonders why it is there but not for very long because she has decided to use it to force her way out so she takes it and goes out to the man at one of the exits and tells him she will shoot him if he doesn't unlock the door and he starts smiling and tells her that that's not necessary because the door isn't locked and he's not going to try and stop her. He told her that he's there to help people get started in a new life once they decide to leave. She drops the gun and sits down laughing at the simplicity of it. She could have avoided all the drama by just saying what she wanted." Import: You've been imprisoned and wanted to force your way out but the people you thought were keeping you in, really wanted to help you; you're happy when you realize you could have avoided the
drama by simply saying what you wanted. Score: plus one.

Both stories are about the theme of Autonomy but the outcomes are drastically different. So the "thematic" similarity is misleading.

The next one is an example of stories with different themes but with the same score. The first is from the pretest of the study, the second is from the "fake good" post-test.

"Agnes thought he loved her. When she let him into her apartment she knew something would go wrong. After he had enough liquor to fill a tank he grabbed her, took advantage of her in drunken rape. She is now totally confused and cries on her bed. She will never find love and will be a depressed old maid." Import: If a person desperately needing to be loved is taken advantage of, he may be damaged for life and never find love. Score: minus two.

"She's an alcoholic who is locked up to break the habit. She has just gone through a nervous tantrum, is bleeding and sore. As she sits there, however, she feels a change within her. She hears the birds whistle and children play and she realizes her problems are solved." Import: A person with an habitual problem may suddenly feel changed within and the problem is solved. Score: minus two.

In this case although the stories, at first sight, appear different both "thematically" and "content-wise" they both get a minus two score.

It seems, therefore, that Arnold's system is more sensitive to true consistency. It doesn't measure "themes" or story content as such but the way the themes are organized in terms of an outcome. Thus changes in the over-all M.I. would reveal changes in the individual's motivational pattern,
not changes in theme or alone. In the first example there is significant improvement in score despite the fact that the stories have the same theme and in the second the score remains low despite the high sounding "Pollyanna" story on the second testing.

Lindzey and Herman (1955) in their study on repeat reliability mention that one of the awkward empirical problems in this kind of study is the contribution of memory to the stories told on the second administration. The consistency between the first and second testings would simply result from the fact that the subject remembered the stories that he told the first time around and repeated them. Thus, one would be measuring consistency of memory rather than the consistency of motivation.

This objection can be answered by continuing the reasoning started above. The issue is not whether the story is similar but whether the outcome to the story is the same. If the person's motivation has changed between testings the change will be evident in a difference in outcome even though the general plot of the story remains the same. If the motivation hasn't changed the subject will get consistent scores even if the stories have different plots. So the awkwardness of which Lindzey and Herman complain is not really empirical but theoretical. It is difficult to get consistency if the TAT is dealt with in terms of themes rather than outcome.

Weisskopf and Dieppa (1951) found that hospitalized veterans could influence their stories in a favorable or unfavorable direction, so that the diagnostic conclusion of TAT interpreters were affected on some but not on all personality dimensions. They used only three pictures, so it left the question open as to whether the subjects could have consistent figures for a full test. In the present study 13 pictures were used. For the results
it seems that it is unlikely that the test can be faked when the normal complement of pictures is used.

Dunlevy (1953) obtained results similar to those of Weisskopf and Dieppa when the examiners were unaware of the purpose of the experiment. When they knew ahead of time that some subjects were trying to make more, and others less favorable impressions, they were able to detect stories told under the different instructions with some success.

In the present study the protocols were scored blindly by two examiners and the results showed no appreciable change between the control and experimental group. The experimental groups could not "fake" well enough to influence the M.I. in a positive direction. Part of the reason for this is that the "Pollyanna" stories described above get negative scores. Thus it happened that some subjects simply retained their already negative scores and others lowered them. This is illustrated by the stories quoted on page 31. The first one an obviously negative import: "If a person desperately needing to be loved is taken advantage of, he may be damaged for life and never find love." The second (fake good) looks positive but it is characteristic import of a low achiever: "A person with an habitual problem may suddenly feel changed within and the problem is solved." Thus, both imports earn a minus two score.

Another example shows how the score can actually get worse when the subject attempts to fake good.

"This was Charles' third rap. He was picked up three days ago for breaking into a retail clothing shop. The other two times the judge had been lenient and the other two times Charles had friends who could get him
out on bail. Charlie needed a fix and he needed it bad. He was already feeling the pain and craving. He was lucky, he thought, that they didn't find that "horse" on him when they brought him in otherwise he'd have went will (sic) a couple of days ago. Charlie was found guilty of breaking and entering and upon arriving at the state pen was found to be a drug addict and is now under medical treatment." Import: If someone breaks the law repeatedly to support an addiction, he'll end up in prison but will get treatment for the addiction. Score: plus two.

"They had beat him up again just because he stands for something the others don't agree with. But he won't buckle now. He can't consider himself open and honest and yield. No. Not ever. He'll get up and then get knocked down again. And though it seems tremendously difficult he'll stand back up, because he's not afraid. He's not afraid to die for what he really believes and he's gonna tell people about it." Import: You're not afraid to tell people what you believe in and die for it if necessary. Score: minus one.

This is a good illustration for two reasons. When the subject took the test under normal conditions he told a seemingly "negative" story. The plot deals with drug addiction, robbery and incarceration. The second, done under conditions of fake good, seems to be much more "positive" for it talks of openness, honesty, determination, and the willingness to face death.

Arnold (1962) found, however, that stories indicating that wrongdoing, ill-intentioned, and imprudent action brings punishment are told by high-achievers. On the contrary, stories dealing with heroics of any kind are
told by low-achievers and, as a result, get negative scores.

The conclusions that seems to follow from this is that the SSA is also relatively impervious to attempts to "fake bad" for apparently negative stories can get a positive score depending on the outcome. This would be an interesting topic for further research.

Davids and Pildner (1958) found that job applicants obtained significantly higher adjustment scores than research subjects on self-reporting devices but not on the TAT and on other projective tests. The findings of the present study tend to confirm their results by showing that it is difficult to make oneself "look" good on the TAT.

Ni (1959) also reported that he was able to detect attempts at concealment for six college students from blind analysis. The problem with his study is that the number of subjects is very small.

Although the goal of the present study was to examine the possibility of over-all "fakability" rather than particular attempts at concealment, these attempts are most evident in the "Pollyanna" stories, which with experience, are fairly easy to detect.

Lastly, Dollin and Reznikoff studied the relationship between the Edwards (1957) Social Desirability Scales (short form) and six TAT cards with 48 psychiatric patients and 56 nursing students. They found no confirmation of an S.D. bias on TAT responses. Subjects with high scores on the S.D. scale did not consistently show high S.D. tendency in their stories.

It is practically impossible to relate the present findings to those of Dolling and Reznikoff since their method and scoring procedures are
quite different from those employed here. The subjects in this study were not tested on their desire to look socially acceptable.
CHAPTER V

CONSIDERATIONS FOR FUTURE RESEARCH

Given the results of the present study, it would be interesting and profitable to pursue two lines of research.

One could study whether, using the SSA, it would be possible to "fake bad" on the TAT. For reasons given above it would seem that the SSA would be equally resistant to attempts to "fake bad".

A second line of investigation would be to compare the SSA to some other personality instrument such as the MMPI or the 16 PF in a normal and "fake good" situation. It is also likely that the "objective" personality test would be easier to fake than the SSA.

Further implications

One of the most difficult areas of research is that of evaluating the effectiveness of psychotherapy. Smith (1973) poses the problem as follows:

"New approaches to psychotherapy generally arise in the clinic rather than in the laboratory. Unlike drugs, which are developed and tested before they are applied, new therapeutic techniques are typically 'invented' and applied before they are researched. The innovative clinician is more in the forefront of new therapeutic approaches than the researcher. However, once a new therapeutic approach is found effective by the clinician in practice, it still needs to be analyzed. Its effectiveness needs to be more rigorously evaluated lest too much be made of startling but scattered successes."

He investigated the effects of such a newly developed group therapy (Tyrell, 1972) on a group of unmarried males in their last two years of
undergraduate studies at a small college. The control group consisted of other students in the same college most in their junior year and a few in their senior year. All subjects were given the MMPI and the 16 PF Personality Factor Questionnaire Form A and the TAT. The TAT was administered, interpreted and scored according to Arnold's Story Sequence Analysis (1962).

He found that when the three measures used to evaluate DC therapy were compared, the SSA of the TAT most clearly distinguished between the two groups at a .001 level of significance.

Linking this with the findings of the present study it would seem that the SSA of the TAT is a reliable instrument for measuring changes occurring due to the intervention of psychotherapy; a) it clearly distinguished between groups receiving therapy and those receiving none; and b) it is not easily faked good indicating that the changes were real changes in motivation.

Another problem to which the findings of this study can be applied is that of assessing the motivation of applicants for certain demanding professions. If one could detect motivational difficulties in individuals prior to their acceptance in long, expensive training programs it would benefit both the applicant and the program's directors and faculty members. It is imperative, then, to find reliable instruments to measure the individual's motivational condition without his being able to look better than he really is. The SSA of the TAT seems to be such an instrument. Arnold (1962) spells out how, for example, this could be done for screening
candidates for religious orders.

Conclusion

It appears, then, that the SSA of the TAT fulfills the plea of Fiske (1971) quoted in the introduction. It is a reliable measuring procedure of a major psychological construct and hence a valuable research tool for exploring motivational dimensions of personality.
REFERENCES


Burhard, M. I. Characteristic differences, determined by TAT sequential analysis, between teachers rated by their pupils at the extremes in teaching efficiency. Doctoral dissertation, Loyola University, Chicago, 1958.


Child, I. L., Frank, Kitty, F., & Storm, T. Self-ratings and TAT: Their relations to each other and childhood background.


Davids, A., & Pildner, H. Comparison of direct and projective methods of personality assessment under different conditions of motivation.

*Psychol. Monograph*, 1958, 72, No. 11 (Whole No. 464).

Dollin, A., & Reznikoff, M. TAT stories and the social desirability scale.


Ni, L. Study on concealment of the subjects in telling stories on TAT pictures. Taiwan, Formosa: National Taiwan University, 1959. Psychological abstracts, 34, 4654.


child development, 1943, 8, No. 1 (Ser. No. 34), 258-301.


Tomkins, S. S. Limits of material obtainable in the single case study by daily administration of the TAT. Psychological bulletin, 1942, 39, 490.


APPROVAL SHEET

The thesis submitted by Roger S. Arnold has been read and approved by members of the Department of Psychology.

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval with reference to content and form.

The thesis is, therefore, accepted in partial fulfillment of the requirements for the degree of Master of Arts.

Cfinere 9th, 1974
DATE

Advisor's Signature

Anne Z. Heilman