THE EFFECT OF VARYING TIME INTERVALS ON THE REPRODUCTION AND RECALL OF RORSCHACH RESPONSES ON RETEST

Thesis For The Degree Of Ph. D.

MICHIGAN STATE COLLEGE

BERTRAM H. SCHNEIDER

1955

This is to certify that the

thesis entitled

The Effect of Varying Time Intervals on the Reproduction and Recall of Rorschach Responses on Retest

presented by

Bertram H. Schneider

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Psychology

M. Ray Denny
Major Professor

Date 5-6-55

THE EFFECT OF VARYING TIME INTERVALS ON THE REPRODUCTION AND RECALL OF RORSCHACH RESPONSES ON RETEST

By

Bertram H. Schneider

AN ABSTRACT

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Sciences in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

Year 1955

Approved M. Ray Denny

THESIS

The purpose of this study was to investigate the consistency of Rorschach results on retest where subjects were not exposed to any treatment other than the systematic varying of the time intervening between the tests. Specifically, it was to determine the differential effects of the passage of time on the persistence, change and recall of Rorschach responses on retest.

Two hypotheses were set forth: 1) performance on the Rorschach in terms of repeated responses on retest remains constant over varied short time intervals; 2) recall of those responses decreases as a function of the length of time between tests.

Sixty patients screened to assure exclusion of those with neuropsychiatric conditions were selected as subjects from a VA general medical and surgical hospital. The subjects were distributed into three groups of twenty, equated for age and intelligence. Each group was retested with the Rorschach after the following approximate time intervals: group I, four hours; group II, two weeks; group III, two months. Following the retest of each subject, his responses were individually read back to him and he was asked if these responses had been given in the initial test. Following this recall procedure a questionnaire, designed to obtain a verbal report of the effect of recall on the retest, was administered to each subject.

Two techniques were used to obtain data to test the hypotheses. The response-comparison technique was a matching

procedure, by means of which each pair of Rorschach protocols were compared for common or consistent responses. The recall technique was a scoring method by means of which accuracy in the identification of retest responses as new or repeated responses was determined. The two techniques yielded seven measures which were tested for significance of differences among the three groups by the t-test.

The results on the whole confirmed both hypotheses. It was found that the measures of consistency devised to test the first hypothesis did not yield significant differences among the three groups, regardless of the length of time elapsed before retest. The measures of recall devised to test the second hypothesis decreased as a function of the length of time between tests. It is concluded on the basis of the results that retest consistency is not to be solely accounted for in terms of recall.

The verbal reports of the subjects are compatible with this finding. Forty per cent of the subjects reported that it was the stimulus properties of the cards that seemed to elicit the same responses on retest. This may be compared to 13% who reported that it was recall that seemed to be of primary importance in eliciting the same responses.

The results also indicated that the percentage of new responses on retest was 25.0 after four hours, 32.2 after two weeks and 29.5 after two months. The verbal reports suggest that ease of concentration, curiosity and desires to be more thorough were some conditions related to the production of new responses.

THE EFFECT OF VARYING TIME INTERVALS ON THE REPRODUCTION AND RECALL OF RORSCHACH RESPONSES ON RETEST

By

Bertram H. Schneider

A THESIS

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Sciences in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

Year 1955

ACKNOWLEDGMENTS

The author is most deeply indebted to Dr. Albert I. Rabin, under whose inspiration and constant encouragement this research was initially undertaken. He also wishes to express his sincere thanks to Dr. M. Ray Denny, who guided the completion of the study, and to other members of the doctoral committee, Drs. Carl Frost and Alfred Dietze for their constructive criticisms and helpful suggestions.

Grateful acknowledgement for valuable assistance is also extended to the psychology staff at the Dearborn VA Hospital, including Drs. John J. Brownfain, Andrew S. Dibner, William A. Alexander, Herbert B. Malos and, in particular, to Dr. Bernard Chodorkoff for his most generous help.

Deep appreciation is also due to the author's family, especially to his wife, Nelda, for her endurance and innumerable contributions throughout the course of the study and to Mrs. Jesse Goldberg for her typing assistance.

And, finally, thanks are accorded to friends and associates, who aided in one way or another, including Drs. Ned Papania and William Brett and to Rosemarie Szilagyi for typing the final manuscript.

TABLE OF CONTENTS

		Page
ACKNOWLEDGEMENTS		iii
TABLE OF CONTENTS		iv
LIST OF TABLES		vi
I.	INTRODUCTION	1
	A. Some Problems of Rorschach Reliability	2
	B. Previous Studies	4
	Split-Half Studies	4
	Parallel Test Studies	6
	Test-Retest Reliability Studies	9
	Experimental Test-Retest Studies	16
	Clinical Test-Retest Studies	23
II.	HYPOTHESES	28
III.	METHODOLOGY	30
	A. Subjects	30
	B. Procedure	31
	C. Methods and Techniques	35
	Response Comparison Technique	35
	Recall Technique	37
	Matching by Judges	40
	Questionnaire	42
	D. Treatment of Data	43
IV.	RESULTS	48
V.	DISCUSSION	60

Page

A. Consistency and Recall 60

B. Some Conditions Related to Consistency 61

C. Some Conditions Related to Inconsistency . . . 62

D. Judges' Verbal Reports 64

E. A Theoretical Interpretation 65

F. Methodology 66

G. Implications for Further Research 66

VI. SUMMARY 68

APPENDIX A - QUESTIONNAIRE 71

APPENDIX B - TABLES 72

APPENDIX C - VERNON'S FORMULAE 86

BIBLIOGRAPHY 87

LIST OF TABLES

		Page
Table I	- Description of Groups	32
" II	- Definitions of Recall Classifications	39
" III	- Summary of Scoring Measures	47
" IV	- Comparison of Groups on Differences in the Reproduction Measure	49
" V	- Comparison of Groups on Differences in the Reproduction-Recall Measure	50
" VI	- Comparison of Groups on Differences in the Total Reproduction Measure	51
" VII	- Comparison of Groups on Differences in the Total Reproduction-Recall Measure	52
"VIII	- Comparison of Groups on Differences in the New Response Measure	54
" IX	- Comparison of Groups on Differences in the New Response Identification Measure	55
" X	- Comparison of Groups on Differences in the Recall Measure	56
" XI	- Comparison of Groups on Mean Percentages In Each Measure	58
" XII	- Comparison of Groups on Replies to Questions 1,3,4,5,6,7,8,9	72
"XIII	- Comparison of Groups on Question #2 of Questionnaire	73
" XIV	- Comparison of Groups on Question #3 of Questionnaire	75
" XV	- Replies of "Yes" to Question #4 of Questionnaire	77
" XVI	- Comparison of Groups on Question #5 of Questionnaire	78

Page

Table XVII - Replies of "Yes" to Question #6 of Questionnaire 80

" XVIII - Comparison of Groups on Question #10(a) of Questionnaire 81

" XIX - Comparison of Groups on Question #10(b) of Questionnaire 83

I. INTRODUCTION

Although the Rorschach test has become firmly established in psychological clinics as a major diagnostic instrument, research with it has never been focused on the specific question of temporal reliability. Most clinical studies on the Rorschach which report changes in performance due to drugs, electroschock, psychotherapy and the like, do not include untreated control groups. Experimental studies generally are of the "testing-the-limits" nature where the standard procedure is altered in order to check on the resistance of the Rorschach to artifacts of the situation. Reliability studies employing the test-retest method have been few for fear that practice or memory effects would mask any instability in the test.

The present study investigates the stability of the Rorschach in a situation where the subjects are not exposed to any treatment other than the systematic varying of the time interval between the initial test and the retest. The use of a temporal dimension permits an evaluation of the differential effects of the passage of time on the persistence, change and recall of Rorschach responses in retest.

A. Some Problems of Rorschach Reliability

Test reliability refers to the consistency with which a test yields information. Two major sources of unreliability are recognized: a) lack of stability in the function which is tested and b) errors in measurement. To illustrate the former one might consider the measurement of an earthworm by means of a foot ruler. The errors in measurement would be minimal whereas fluctuations in the length of the earthworm would give rise to inconsistent results. To illustrate the latter, one might consider the reverse, the measurement of a foot ruler by means of an earthworm. Here the function is perfectly stable, whereas the measuring instrument itself is faulty.

This problem is particularly applicable to psychometrics where both fluctuations in function and errors in measurement are common. If one attempts to establish test reliability by comparing the results of two test administrations separated in time (test-retest), both sources of error are operating. As an alternative the split half technique has come into general usage. This method makes use of the comparison of equivalent halves of a test (first half vs. second half; odd items vs. even items, etc.). The obvious advantage here is that an assumption of function stability does not have to be made in order to assess errors of measurement. However, where changes in the function over a period of time is the subject of investigation, the retest technique is essential. In this case the function must be

considered stable over a short period of time and the errors of measurement then estimated between the initial and the repeated tests.

The test-retest approach to reliability has long borne the stigma of practice or recall as a source of error that gives a false appearance of consistency. Still, surprisingly enough, although the concept of reliability has been known since 1904 and the literature dealing with test reliability has increased to sizeable proportions, there is little evidence that this assumption has ever been directly tested.

Jackson (24) reports a study investigating the effect of varying the time between tests on reliability by using the Revised Beta Examination. The time intervals were one-half day, one day, three days, one week and five weeks. He found no general pattern evident in the results and he was unable from a statistical point of view to determine the net effects of practice. It did appear, however, that changes were not related to the length of time elapsed between tests. This could be explained on the basis that Jackson tested a relatively stable function (intelligence) and hence errors of measurement would remain constant over the various intervals of time. If memory effects were significant in sustaining an impression of stability, changes should have increased with the passage of time.

With the advent of personality tests, the problem of determining consistency through retest lead to a dilemma. On the one hand is the argument that lack of change in the

retest may be accounted for by memory or practice. On the other hand is the possibility that this lack of change is due to the reflection of stable features in the personality and that any changes are a consequence of actual psychological changes in the subject.

Opposing points of view on this problem have been represented among Rorschach investigators. Thornton and Guilford (49) state that the importance of the memory factor precludes a repetition of the same test series. Consequently they argue that the split-half method is the only one possible with the Rorschach technique for reliability studies. However, Piotrowski (37) insists that the split-half technique is unfeasible because of the unitary nature of the test which makes the direct comparison between parts impossible. He further claims that there are no practice or memory effects, because there is no conscious effort. He feels that rather than be called mere repetitions, repeated percepts should be considered as representing stable personality trends typical for the time elapsed between examinations.

B. Previous Studies

Split-half studies. The pioneer study on Rorschach reliability can be attributed to Vernon (51). There were two earlier test-retest type studies by Mira (34) and Wertham and Bleuler (53) which will be described later. Neither of these two studies presented statistical findings and Vernon's comment on them was that owing to the effect of memory factors, any "repeat correlation coefficients" would be spurious. He, therefore, decided to use the split-half method by considering the test as consisting of two parallel series of five

blots each. He correlated the responses from one series consisting of cards I, III, V, VI and X with cards II, IV, VII, VIII and IX. The Spearman-Brown formula was used to correct for the reduced length of the test. The results yielded an average correlation of .54 which Vernon considered unsatisfactory. The one exception was a correlation of .91 for the number of responses. His conclusion was that if the test is to have any claim to "objective validity" it must be modified in order to achieve a minimum reliability level of from .70 to .80.

Hertz (20) followed Vernon's study with one also using the split-half method in which she divided the test into odd and even numbered cards. She reports finding an average correlation coefficient of .829 as compared to Vernon's average of .54. She explained the higher reliability on the basis of more adequate standardization of the testing procedure and increased objectivity of the scoring. As a consequence of her findings Hertz maintained that the test factors were reliable and that hence personality traits in terms of the inter-relationships of the Rorschach factors tend to be consistent and follow a stable pattern.

A third investigation using the split-half technique was reported by Thornton and Guilford (49). They singled out the "Erlebnistypus" scores for a reliability study. The results were somewhat inconclusive. They stated that, under favorable but unknown conditions, reliability could be demonstrated for the M and C scores. This study touched off

a blast by Piotrowski against "atomistic" studies carried out to avoid the memory factor. Shortly after this a general reaction against split-half studies set in among Rorschach workers. Hertz (21), herself, signalled the end of this method by pointing to the "global nature of the test" which excludes working with variables abstracted from the whole.

A further re-evaluation by objectors to the split-half method indicates its unsuitability for the following reasons: 1) There are unique stimulus values in the individual cards. Hence each card elicits varying frequencies of the responses that are summarized in the various scoring categories. 2) The split-half method assumes a relatively constant test performance throughout the test situation which is not met in practice. The cards have a sequential and mutual relation considered of interpretive importance which is not only of necessity ignored in split-half studies but also is contrary to the assumption of an unfluctuating function. 3) On statistical grounds Cronbach (10) points to the unfeasibility of using the Spearman-Brown correction, where ratio scores, such as those found in the Rorschach, vary with productivity. 4) The fact that five of the ten cards are colored precludes the equality of the stimulus values of the split-halves.

Parallel test studies. A different approach to the problem of the memory factor through the use of an alternate

or parallel form of the inkblots has often been advocated. Rorschach, himself, indicated what he felt was the necessity of such a series, saying, "If the test is to be repeated with the same plates, conscious or unconscious memory may warp the results. Analogous series of plates... are necessary for these situations," (43). Of the number of alternate forms which have been devised, the Behn series has recently received the most attention.

Swift (48) carried out one of the first studies directed at measuring the degree of correspondence between the two test series. Using the Rorschach as the initial test, she retested preschool children with the Behn set after a seven day interval. Correlations of the scoring categories from both tests yielded some high coefficients, but a number of low ones led Swift to conclude that the Behn was not sufficiently comparable to the Rorschach.

Eichler (12) reports a more recent study involving the Behn series. He used three groups, retested after approximately three weeks, the first group receiving the Behn followed by the Rorschach, the second receiving the tests in the reverse order, while the third served as a comparison group with the Rorschach followed by the Rorschach. His findings were similar to that of Swift's in that while reliability coefficients from the Rorschach retest as compared with the Behn retest were satisfactory in some

respects, consistent differences in other respects indicated the two tests were not sufficiently alike for use on the individual level.

Singer (46) used a more "global" approach to the Rorschach-Behn comparison. He gave the protocols from ten subjects who had been given the Behn and the Rorschach to six judges with instructions to pair them. Although the matchings were better than chance, Singer concluded that the Behn failed to meet standards of reliability demanded for individual prediction.

The consensus of these findings is that the Behn is not an entirely satisfactory alternate form, especially for use in the individual case. If one examines alternate forms from a theoretical point of view, its disadvantages become apparent. While it is true that memory effects are eliminated as a matter of concern for reliability, the alternate form contributes another source of error, i.e. the extent to which it falls short of being equivalent to the standard test. This lack of equivalence is especially prominent in projective techniques where the unstructured material is so difficult to duplicate. The more accurately the projective test is duplicated the closer the alternate form comes to being identical with the original test. It then becomes more than a retest and subject to the possible effects of repetition.

Test-retest reliability studies. The persistent concern with the memory factor in retesting has led to two studies which attempt to control for it in unique ways. In one study, Kelley, Margulies and Barrera (26) chose patients after electroshock who had amnesia for the initial test which immediately preceded the shock. Twelve of these patients who were free from confusion were retested two hours later. The authors describe the changes as minor with few shifts of more than one response in the variables. The general personality picture of each appeared unchanged, although no statistical verification was reported. The fact, however, that there were some changes may be due, as Rabin (39) points out, to cerebral changes concomitant with electro-shock treatment.

Griffith (18) made use of patients with Korsakoff's syndrome as a means of ruling out memory factors, since gross memory defects are an integral part of this disease entity. He found four patients who appeared to have no recall for the test upon the retest 24 hours later. A comparison of each pair of test-retest protocols showed similar features which reliably characterized each individual. Full statistical treatment of the results was precluded by the small number of subjects.

Despite the ingenuity of such studies as these, the results cannot be wholly definitive on the problem of memory. The subjects used are seldom encountered in practical

clinical experience. Also, the actual effect of treatment or disease conditions on test performance must be discovered through further research. To put it more simply, the more practical question is the extent to which memory actual affects the test when it is repeated, since it is not the usual case for a subject's memory to be blotted out in the intervening time.

In spite of the contention that recall would invalidate reliability studies using test-retest, such studies have been attempted. One of the earliest studies of this nature was carried out by Mira, (as reported by Vernon) (51). He administered the retest to a group of subjects two weeks after the initial test. Some consistency of responses was observed in some subjects. Mira considered the degree of consistency as an index of the stability of the individual. Since he did not present any statistical evidence of his results, the study is more of historical interest than of significance for this research.

Several studies involving children as subjects have been reported. Kerr (27) repeated the test with fifty elementary school children who had been first tested the previous year. She compared the first and second tests in terms of correlation coefficients for several scoring categories, which ranged from .001 to .74. Color determinants fell in the lower ranges. Kerr explained this finding on the basis of the affect represented by color responses.

It should not be surprising that color varies so greatly since the emotional state of the subject is similarly inconstant. On the other hand, the number of whole responses, which yielded the highest correlation coefficient, was said to be expected, since Rorschach indicated that it was highly correlated with intelligence. It should be noted here that Kerr's use of correlation coefficients of the scoring categories as the sole statistical criteria of reliability without considering the configurations of the Rorschach factors is only a partial approach to the determination of reliability.

A similar study, but with a preschool population of 55 subjects, was reported by Ford (13). While admitting the incomplete nature of a statistical approach of the type employed by Kerr, she could see no way out but to express the results in Pearson product-moment coefficients. The reliabilities ranged from .38 to .86 with each coefficient indicating a significant relationship between test and retest determinants. She stated that although these findings are not high, they can be considered as fairly satisfactory especially since the final synthesis depends on the balance and interrelationship of all the determinants.

Troup (50) provides an interesting and well designed study from a methodological standpoint, illustrating a means of sidestepping the limitations of the piecemeal correlational procedures. In this study, six judges were asked to match

two Rorschach psychograms taken six months apart for each individual of ten pairs of twins of grammar school age. Comparing ten pairs of Rorschach psychograms at a time, three judges achieved 100% matchings, one, 90%, and two, 80%. Using a formula developed by Vernon (51), this yielded a contingency coefficient of .94. A chi square test of significance indicated the chance expectation of this figure was less than .001. Troup's conclusion was that the "degree of reliability based on the consideration of the total personality picture appears significantly greater than estimates based upon...correlation coefficients of the separate categories."

The previously mentioned study by Swift (48) is also pertinent here and will be described in greater detail since it is particularly relevant to this investigation. At varying time intervals Swift tested preschool children under four conditions: 1) A Rorschach test-retest with a median interval of 30 days using 41 children. This interval was chosen in the belief that the memory factor would be minimized while developmental factors would remain constant. 2) A Rorschach test-retest with a median interval of 14 days. The Behn series was interpolated on the seventh day. The subjects numbered 49 of whom 19 participated in condition 1, given the previous year. 3) The Behn series of condition 2 with the 7-day interval. 4) A Rorschach test-retest with a ten month interval. This latter group included 20 subjects, all of whom had been used in conditions 1 or 2. The use of

the same subjects in the various conditions were justified by Swift on the basis that they apparently did not recognize the test from the previous year.

The reliabilities of the various conditions involving Rorschach test-retest were reported in terms of product-moment correlations for those scoring categories which presented a continuous distribution of scores. For the 14 day interval the reliabilities ranged from .59 to .84 (corrected for attenuation) with nine of twelve categories over .70. The two month interval yielded a range of .15 to .87 while the ten month interval produced a range of .08 to .86 with all but one over .50. It would appear that the highest reliabilities are attained when the interviewing period is brief. However, it would be hazardous to generalize from data collected under the conditions reported. It will be recalled that some subjects were reused in the various testing situations. It is even likely that some of the subjects tested in condition 4 had already been tested in all of the other conditions, making a total of six separate test administrations for them. Furthermore, the statistical analysis by the standard correlational methods is subject to the same criticism applied to those studies mentioned above, i.e. ignoring the patterns of scores.

Another feature of Swift's study was an attempt to determine the extent of repetition in the retest. An analysis of the responses was made to determine the average

percentage of identical responses in the initial and succeeding tests. For the 14 day interval Swift found the average percentage of responses given in the retest, which were identical with those of the original test, to be 57% with a range of 0% to 90%. The corresponding percentage for the 30 day interval was 51% with a range of 8% to 100%. Here it would seem that there is little variation in repetition from one time interval to the other. One should not assume, however, that an identical response is produced in retest because of recall. Although Swift cautions against considering these percentages as an index of memory factors, it has been so represented by Ainsworth (1).

A final study which might be added to this section is reported by Holzberg and Wexler (22). Working with 20 chronic schizophrenic subjects, they sought to determine the reliabilities of the Rorschach when used on a population clinically defined as "unpredictable". The underlying assumption was that the retest might be significantly changed without a corresponding change in the clinical picture because of the "unpredictability" of the schizophrenic. A three week period intervened between the two testings. Statistical procedures used were correlations of scoring categories and tests of differences between means. On the whole the results revealed significant correlations between means. Since some qualitative variation was observed in a few pairs

of protocols, the question was raised whether the statistical analysis masked important differences which might influence clinical judgement in identifying personality structure from test to retest. To answer this question tabulations of the data were submitted to two trained Rorschach workers. Matchings of pairs of tests were significant at less than the 1% level of chance occurrence. The authors concluded that "unpredictability" was not apparent in test-retest performance with a schizophrenic population which has become stabilized in terms of chronicity.

The studies listed above represent the bulk of the literature dealing exclusively with the retest criterion of reliability. A summarizing review of these studies reveals numerous shortcomings. The limitation of subject populations to children and schizophrenics does not provide information on the most representative individual of the population at large, the normal adult. Some data related to normal adults are available in the experimental studies, which are to be reviewed next. Choice of time intervals in the preceeding studies has been largely arbitrary except in Swift's study (48) where the methodology was faulty. The statistical treatment of data in the earlier studies was a perseveration of the method used in split-half studies, dealing with the separate scoring categories instead of the interrelated patterns of scores which constitute the major interpretive

unit. The question of recall and practice effects has received little attention, again with the exception of Swift. However, the design of her study did not involve assessing the net effect of memory in the reproduction of responses.

Experimental test-retest studies. A number of experimental studies, using the retest technique, have been designed to take account of extraneous factors on Rorschach performance or, as one writer (1) put it, to "test the limits" of reliability. The usual method in these studies is to alter the standard test conditions in order to determine the possible effects on the stability of the test.

The historical forerunners of these studies were carried out by Fosberg (14,15) in an attempt to check the resistance of the Rorschach to "faking". In a preliminary investigation (14) he used two subjects and in a latter one, 50 subjects. The design of both studies involved the administration of the test four times to the same subjects; first, with standard instructions; second, with instructions to make the best possible impression; third, with instructions to make the worst possible impression; fourth, with standard instructions again. The retest intervals with a range of from 0 to 700 days were not held constant. In the first study Fosberg used a chi square to show that the psychograms for each person corresponded. In the second, a correlational

technique was used. The results of both studies led him to conclude that the Rorschach could not be manipulated with the deliberate intention of presenting oneself in a favorable or unfavorable light. He also stated that the time interval apparently had no effect on the reliability coefficients in the second study. However, Cronbach (10) indicates that the statistical procedures used throughout both studies were "entirely unsound" and hence all the conclusions are open to question.

Carp and Shavzin (8) attempted to verify Fosberg's findings by further testing the susceptibility of the Rorschach to falsification. They tested 20 subjects twice, three weeks apart, with instructions to make a good impression on one test and a bad impression on the other. A test comparison of each category in the two tests showed no group differences, although there were differences on the individual level. When the two tests of each subject were compared as units by chi square tests, it was found that for four subjects the probabilities were less than .10 that the two distributions came from the same population. They considered the results as refuting Fosberg's statement that the Rorschach resists <u>all</u> attempts at manipulation by the subject, since their evidence shows that some subjects can alter their personality picture as reflected by the Rorschach.

Another study employing altered instructions in the experimental conditions was reported by Hutt, Gibby, Milton

and Pottharst (23). They investigated the extent of modification in specific scores on the Rorschach with specific instructions to alter these scores. Four groups of college subjects originally tested under standard instructions were retested two weeks later, each group under one of the following instructional conditions: 1) to pay particular attention to segmented areas of the blot; 2) to find as many human movement responses as possible; 3) to give only good form, but to combine color and form, and in addition human movement responses; 4) to report everything they saw (standard instructions). This fourth group served as a comparison group. The results of the experimental groups showed a general shift in the direction indicated by the instructions. The conclusion of the authors was that the variables in question were unstable as a result of the test conditions for their non-psychiatric population. As for the control group, which is of interest as a study of test-retest reliability, the findings showed great variability in correlation coefficients of some scoring categories. Surprisingly enough, the correlation coefficients for the control group were lower than those of the experimental group. The authors suggest that this instability, both in the control and experimental population, may be a result of the lack of rigidity in the normal individual. One may well wonder, however, to what extent the college population is representative of the general population.

The above studies all are related in that their attempt to show the influence of "set", suggesting that determining factors in perception imposed by the examiner through the instructions are equivalent to the inner determining factors which the subject brings to a standard testing situation. This, of course, may not be the case. Evidence by Norman et al. (36) and Rabin et al. (40) indicates that subjects exposed to more indirect set-inducing experiences do not generally show the effects thereof. However the implication of the studies using altered instructions is that the instructions must be standardized in order to produce comparable data. Differences in instructions may be responsible for the variance contributed to test scores by examiner differences as noted by Baughman (3). Certainly this must be a factor to be considered in not only diagnostic testing but also any study employing the Rorschach test.

Other studies using the test-retest procedure have been designed to examine the influence of situational factors on the test. Kimble (28) studied the effect of a social setting by administering the test twice, one under standard conditions and another time in a college cafeteria with at least two other people present. Fourteen college students served as subjects. The time elapsing between tests was between one and two weeks. The social situation reportedly elicited significantly greater color responses. Kimble concluded that the friendly, intimate atmosphere evoked in the social setting was responsible for the increase in color,

representing a response to pleasant stimuli.

Another study of this nature was made by Lord (33) who sought to assess the relative effects of retesting, of experimental alteration of the emotional climate in the test situation and of the examiner's personality. She used 36 college students tested under each of three conditions: 1) a standard test situation; 2) following other tests designed to make the subject feel rejected and a failure; and 3) following conditions designed to make him feel accepted and successful. Each examination was administered by a different one of three examiners. The results indicated that retesting effects were the least important. The test atmosphere produced more responsive (R), more imaginative (M), and less stereotyped (A%) results in the approving situation while tendencies in the opposite direction were observed in the disapproving situation. The most prominent effects were those of the three examiners themselves.

These studies also may be evaluated in terms of another possible source of unreliability, differences in subject-examiner relationships. It also indicates the advisability for the individual examiner in diagnostic work to appraise himself of his personal effect on a patient's performance.

The final pair ofexperimental studies using the retest technique are those which alter the test procedure or the stimulus itself. Rabin and Sanderson (41) investigated the

effects of reversing the order of presentation of the Rorschach cards, thus testing the importance of the "temporal gestalt" in producing shock and its associated phenomena (delayed reaction time, decreased productivity, etc.). Two groups of student nurses, 17 in each group, served as subjects. The design was counterbalanced so that in one group the normal order of presentation was followed in the initial test and the reverse order some two months later in the retest, while the converse was true of the other group. The results indicated a few changes as a consequence of reversing the order. Some cards appeared to elicit fewer responses and to prolong response time, regardless of the order of presentation. The authors concluded that shock may be a consequence of the "greater difficulty of some cards and lesser potentiality of others to evoke responses". A further observation was the high stability of the experience balance from test to retest, an interesting finding in view of the susceptibility of this ratio to change under the artificially contrived test conditions of social setting (28) and emotional atmosphere (33).

A study by Allen, Manne and Stiff (2), primarily concerned with the influence of color on retest reliability, is of methodological importance for this study. Two groups of college students totaling 25 were tested with a standard series of cards in one group and with a set from which the color was removed in the other. They were retested six

weeks later with the standard and achromatic sets alternated between the two groups. The responses on each card of both tests were compared for "consistency", defined as "the reappearance of a response in the retest protocol". The mean percent of consistency for the colored cards in the standard set was 30.4% while the percentage for the same cards in the achromatic set was 27%. The same statistic applied to the non-color cards of both sets, standard and achromatic, was 34.6% and 30.6% respectively. Since neither of the differences was statistically significant, the conclusion was that the presence or absence of color seems to have no influence on the degree of consistency. The question that might be asked here is to what extent is the percentage of consistency influenced by memory?

A general review of these experimental studies is in order here. The majority of these studies report shifts in various Rorschach factors under the impact of artificial manipulations of test conditions. They serve to emphasize the importance of maintaining rigorously standardized conditions of administration in order to minimize errors of measurement. At the same time it should be kept in mind that the conditions with which the studies deal are not those encountered in the practical experience of the clinical psychologist. As contrasted with the test-retest reliability studies which largely dealt with children, the experimental studies

generally have depended on student populations for subjects. Hence, in neither case are the results representative of the wider population. As in the reliability studies, the time intervening between the two test administrations has not received attention.

Clinical test-retest studies. The preceeding studies may be described as test-oriented in that there is an underlying assumption of personality (function) stability and it is the test (measurement) itself that is studied. The studies which follow can be considered subject-oriented, since it is the conditions of the subjects themselves that are varied, the implicit assumption being that the test is stable. Because of this premise of test reliability, these studies rarely include comparison groups. Actually such investigations are more related to tests of validity than reliability. Their significance for this particular investigation is that they have a bearing on the important question in the theory of reliability of the stability of the function tested, in this case the personality.

There is a very large group of clinical studies using test-retest, too numerous to describe in detail. Only representative studies or those most relevant to the present investigation will be cited more than briefly.

The historical precedent for these studies is the one by Wertham and Bleuler (53). They administered the Rorschach as a method of investigating differences in the

reactions of individuals under normal conditions and when under the influence of the drug, mescaline. They found comparatively close agreement in the two sets of responses. The slight differences occurring did not materially affect the interpretation as a whole. However, statistical verification of the results was not demonstrated.

Many other studies have been reported on the use of Rorschach retests as a means of evaluating physical treatment through drugs or electroshock treatment: (insulin) Piotrowski (37), Halpern (19), Kisker (29), Bradway (6), Graham (17), Beck (4); (sodium amytal) Kelley and Levine (25); (Metrazol) Kisker (29); (electroshock) Bradway (6), Kelley, Margulies and Barrera (26).

Both Halpern (19) and Kisker (29) report that their schizophrenic populations after treatment seem to retain psychotic patterns in their retests along with evidence of improvement. Beck (4) states that after treatment the "main outlines of the Rorschach pattern are always recognizable belonging to the same individuals" but there are also "changes in important features". He emphasizes that such changes are at the periphery and not the core of the personality.

Some studies have used the Rorschach as a measure of psychotherapeutic changes: Brosin and Fromm (7), Krout, Krout and Dubin (30), Rioch (42), Brosin and Fromm (7)

report that some Rorschach factors appear to remain stable over a course of psychotherapy. Rioch's study (42) included 36 patients undergoing analytically oriented intense psychotherapy. A comparison of the "before and after" protocols indicated some changes representing improvement, but on the whole, pairs of tests were more alike than different.

Another group of Rorschach retest studies that lies somewhat in the border zone between clinical and experimental studies are those that involve hypnosis as a method of artificially altering the emotional state of the subject: Sarbin (44), Levine, Grassi and Gerson (32), Lane (31), Counts and Mensh (9). These studies are reported here because they are of the subject-oriented type according to the dichotomy indicated previously.

The general pattern of these studies has been to retest a small number of subjects (one in several cases) under a variety of suggestions designed to vary emotional states. The most frequent impression reported here is of a stable core of personality running through all the records of an individual along with changes consistent with the hypnotic suggestions. Although the findings are not to be taken as conclusive because of the small number of subjects, they are congruent with findings in the other clinical studies, leading to a conclusion of changes occurring in response to treatment within a matrix of a stable pattern of personality.

As for the significance of the clinical studies for this research, they appear to justify an assumption of relative function stability. How relative is this function stability?

The term "personality structure" would seem to imply by definition a continuous, consistent core of organization. Yet at the same time personality theory takes account of the changing nature of this organization, in terms of its dynamic aspects. Over the course of a lifetime an individual's personality would reflect the influences of major life experiences. If the Rorschach validly taps the personality structure, one should expect retesting after long periods of time to show both peripheral fluctuations along with a stable nucleus of personality. Such is the "relative" function stability found in the literature cited above. Conversely it would be expected that stability would be maximum in retests over short periods of time, where there is no reason to anticipate major personality changes. Hence, a comparison of Rorschach test and retest results over varying short time intervals should indicate consistency regardless of the time involved.

This consistency should be accountable for in terms other than pure memory. Studies of retention beginning with Ebbinghaus in 1885 have indicated that memory varies as a function of time. The typical retention curve has as its main characteristic a rapid decline immediately after learning and a gradual leveling off as time advances. In terms of

the Rorschach this suggests that recall for the first test should decline as the intervening time between tests is prolonged. Consequently, in Rorschach testing, retention over short periods of time should be less pronounced than consistency.

II. HYPOTHESES

On the basis of the foregoing discussion the following hypotheses were formulated:

I. Performance on the Rorschach test remains constant over varied short time intervals between test and retest.

   a. Exact response reporductions (identical responses common to both the initial test and the retest) will not vary significantly as a function of the length of time between tests.

   b. Total response reproductions (both identical responses and responses with changes in locations upon retest) will not vary significantly as a function of the length of time between tests.

   c. New responses (responses appearing for the first time in the retest) will not vary significantly as a function of the length of time between tests.

   d. Successful matchings of test and retest from the same individuals will be independent of the length of time between tests.

II. The reflection of memory for the initial test in the retest will vary as a function of the length of time between tests.

   a. The correct recall of exact response reproductions

will decrease as a function of the length of time between tests.

b. The correct recall of all response reproductions (both identical responses and responses with changes in location upon retest) will decrease as a function of the length of time between tests.

c. The correct identification of new responses will decrease as a function of the length of time between tests.

d. The correct recall of all response reproductions and the correct identification of original responses will decrease as a function of the length of time between tests.

III. METHODOLOGY

A. Subjects

Sixty subjects were used in this study. They were selected from male patients at the VA General Medical and Surgical Hospital in Dearborn, Michigan. The subjects showed no evidence of a neuropsychiatric condition and none had taken the Rorschach previously. Absence of neuropsychiatric disorders was established on the basis of hospital records including reasons for admission and ward behavior, as observed by nurses and physicians. Thirty of the patients were admitted to the hospital for treatment of pulmonary tuberculosis while the other thirty patients were admitted for surgical, orthopedic or general medical treatment.

The sixty patients were drawn from a total of 83, who were originally tested for the purposes of this research. The 23 patients, who were not included in the population of this study, were eliminated for the following reasons: 1) discharge from the hospital before the stated interval between tests had elapsed - 19 patients; 2) psychiatric referral by the ward physician or the observation of a psychiatric condition after the initial test - three patients; 3) a pronounced lack of motivation by one patient upon

retest, manifested by excessive rejection of cards - one patient.

The subjects were distributed into three groups of twenty each, differentiated on the basis of the time intervening between the initial test and the retest. The three groups were retested after the following approximate time intervals: Group I, four hours; Group II, two weeks; Group III, two months. Half of each group (ten subjects) was under treatment for pulmonary tuberculosis, while the other half was being treated for non-tubercular conditions. Tubercular patients were included in the sample because the relatively short length of hospitalization of non-tubercular patients precluded retesting sufficient subjects for Group III in a reasonable period of time. Hence each group was devised so as to be comprised of an equal number of tubercular and non-tubercular patients as a balancing measure.

The three groups were equated for intelligence, as measured by the vocabulary scale of the Wechsler-Bellevue Form I, and for age. Table I is a comparison of the three groups on intelligence and age. In addition to the 83 subjects initially tested with the Rorschach, approximately twenty more were administered the vocabulary scale and excluded from the sample because their intelligence level was not comparable to that of Group III.

B. Procedure

Since there is evidence that inter-examiner differences contribute much of the variance to Rorschach results (3),

TABLE I

DESCRIPTION OF GROUPS

	Mean	Range	S. D.
Age			
Group I	29.4	21 - 40	5.65
" II	30.3	21 - 48	7.03
" III	29.3	22 - 39	8.75
Wechsler Bellevue Vocabulary[1]			
Group I	24.65	15 - 35	5.84
" II	23.6	17 - 34	4.96
" III	24.2	18 - 33½	6.52
Time elapsed between tests			
Group I (hours)[2]	3:59	3:10 - 5:0	33.1[3]
" II (days)	15.3	12 - 18	1.77
" III (days)	65.1	55 - 83	5.84

1. Raw Score.

2. Time determined from completion of inquiry of initial test to beginning of retest.

3. Minutes.

(33), all subjects were individually tested and retested by the author. In order to avoid a research "set" the tests were presented as part of the "routine examinations" administered by the psychology personnel to all hospital patients. Consequently it may be assumed that the test administration elicited responses typical of the normal clinical testing situation. Personal observation and evidence derived from a structured verbal report following the retest tend to confirm this.

Before the administration of the initial Rorschach test the vocabulary scale of the Wechsler-Bellevue intelligence test was administered to each patient. As previously mentioned, approximately a score of patients were eliminated on this basis for the following reasons: 1) too great a deviation from the mean vocabulary score of Group III, the first group tested; 2) a bilingual background which rendered the vocabulary score unreliable as a measure of intelligence.

Following the administration of the vocabulary scale the standard Rorschach test was administered. Based on Beck's prescription (5) the instructions for the first test were as follows:

> "You will be given a series of ten cards one at a time. On the cards are designs made up out of ink blots. Look at each card and tell me what you see on each card or anything that might be represented there. Look at each card as long as you like; only <u>be sure to tell me everything you see on the card</u> as you look at it. When you are finished with a card, give it to me as a sign that you are through with it."

The Rorschach test was readministered after the stated interval depending on the group to which the patient was assigned. The instructions for the retest closely followed the earlier instructions but were designed to take account of the obvious fact that the subject had already been exposed to the test. Another sentence was introduced before the last sentence of the above instructions and given as follows: "Now you might see the same things as before and perhaps something different, but as before tell me everything you see on each card as you look at it."

Upon completion of the retest, the responses of the retest were read back one by one to the subject and he was asked if he had given these responses in the first test and if they were given in the same location. The instructions were as follows:

> "I am going to give you each card again and repeat the things you said you saw this time. I will also show you where you saw them. Please tell me whether you mentioned seeing them in the same place the first time you took this test. You may answer 'Yes', 'No' or 'Not sure'."

After completion of this recall procedure, the verbal report mentioned above was obtained in the form of responses to a questionnaire (see Appendix A). It was anticipated that the retest nature of the project might be disclosed by retested patients to other patients in their wards, who were awaiting retests. For this reason the purpose of this investigation was explained to each patient following the

administration of the questionnaire and his strict confidence concerning the project was requested. A check on the extent to which this confidence was kept was provided by question #1 of the verbal report. The response to this question indicated that more than 95% of the patients did not anticipate being retested when called for the second examination. Of those (two patients) who reported some expectation of being retested, the evidence indicates that they were not so informed by other patients but rather surmised this without full certainty. Hence it may be assumed that there was no deliberate recall practice of the first test prior to the retest.

C. Methods and Techniques

Response-comparison technique. The principal method used for obtaining data to test the first major hypothesis was a matching technique, by means of which the pairs of Rorschach protocols were compared for common responses. The use of this technique has been cited above in the review of previous studies (2), (48). The advantages of such an approach are: 1) it minimizes the extent of examiner influence since matchings are made solely on the basis of the free association; 2) it provides a basis for a study of the differential effect of memory; 3) it offers a more holistic approach to reliability than correlational techniques based on separate scoring categories; 4) it is possibly more

relevant in view of a current trend toward increasing use of content from the force association in interpretation.

The following are definitions of the various categories developed for this study to classify the comparison of responses from a pair of tests obtained from the same person:

> Response reproduction (r) is defined as a response on the retest which is elicited from the identical area of the same card, and, which has the same specific (unelaborated) content as a response in the first test.
>
> Response reproduction, location changed (rl) is defined as a response in the retest which has the same specific content, and, which is elicited from the same card as a response in the initial test but with a change in the location of the area on the card that initially elicited the response.
>
> A new response (nr) is defined as a response in the retest which did not appear on the same card in the initial test.

The chief criticism that might be directed at this comparison technique involves the question of reliability. The major source of unreliability in classifying responses in the retest according to the above outline appeared to be in the category, response reproduction, location changed. The problem here was whether a minor change in the location of a response reproduction automatically made it inconsistent with the response in the initial test and hence classifiable as a response reproduction, location changed. Experience indicated that minor differences in the location were generally due to an incomplete definition of the location

of the response by the subject on one of the tests. Furthermore the interpretive nature of two identical responses with minor differences in location would be exactly the same. Accordingly a set of criteria was listed as an objective guide for discounting minor differences of location. Thus, a response in the retest was considered a response reproduction despite slight differences of location if the following conditions were met: 1) form quality remaining constant, e.g. F/ still F/; 2) location scoring remaining constant, e.g. D still D; 3) emphasis on primary content category remaining constant, e.g. A still A; 4) scoring of a "popular" remaining constant even if location scoring shifted, e.g. P still P, despite change from D scoring to W scoring or the reverse.

A test of the reliability of the response-comparison technique was performed by submitting a sample of ten pairs of records chosen at random with the above definitions and special criteria to another psychologist for matching and classification. The percent agreement with the original scoring was 90. This was considered a satisfactory level of scoring reliability.

Recall technique. The recall technique was designed to obtain data to test the second major hypothesis. As described in the previous section[1], this method involved reading back to the subject his retest responses and asking

[1]See page 34 for description of procedure and instructions.

him if he had given these responses in the first test. One of three possible answers, "yes", "no" or "not sure", was scored for each response. This scoring combined with the three classes derived from the response comparison provided three subclasses of each category of the response matching. The following are definitions of each classification employed in the scoring of the recall technique: (See Table II for a summary of these classifications)

Correctly recalled reproduction (cr) is defined as a reply of "yes" by a subject to a response reproduction (r), when he was asked by the examiner, repeating the responses of the retest, if he mentioned seeing it and in the same place in the initial test.
Questioned reproduction (qr) is defined as a reply of "not sure" to a response reproduction.
Incorrectly recalled reproduction (ir) is defined as a reply of "no" to a response reproduction.

Correctly recalled reproduction, location changed (crl) is defined as a reply of "no" to a response reproduction, location changed, (rl).
Questioned reproduction, location changed (qrl) is defined as a reply of "not sure" to a response reproduction, location changed.
Incorrectly recalled reproduction, location changed (irl) is defined as a reply of "yes" to a response reproduction, location changed.

Correctly identified new response (cnr) is defined as a reply of "no" to a new response (nr).
Questioned new response (qnr) is defined as a reply of "not sure" to a new response.
Incorrectly identified new response (inr) is defined as a reply of "yes" to a new response.

Two additional classifications were found necessary to categorize "rejections" of cards. These are defined as follows:

Correctly recalled rejection (crj) is defined as a reply of "yes" by a subject when asked if he failed to see anything the first time on a card which was

TABLE II

DEFINITIONS OF RECALL CLASSIFICATIONS

Code	Classification	Subject's reply-	Reply should be
cr	Correctly recalled reproduction	Yes	Yes
qr	Questioned reproduction	?	Yes
ir	Incorrectly recalled reproduction	No	Yes
crl	Correctly recalled reproduction, location changed	No	No
qrl	Questioned reproduction, location changed	?	No
irl	Incorrectly recalled reproduction, location changed	Yes	No
cnr	Correctly identified new response	No	No
qnr	Questioned new response	?	No
inr	Incorrectly recalled new response	Yes	No
crj	Correctly identified rejection	Yes	Yes
		No	No
irj	Incorrectly identified rejection	Yes	No
		No	Yes

rejected both in the initial test and in the retest and as a reply of "no" if the card was not rejected in the initial test.

Incorrectly recalled rejection (irj) is defined as a reply of "no" by a subject to the same question as above when the card was rejected in the initial test and as a reply of "yes" when the card was not rejected in the initial test.

The scoring of classifications in the recall technique, based as it was on the three possible replies of a subject, was entirely mechanical and should not raise any question of reliability.

Matching by judges. A supplementary method used to obtain data to test the first major hypothesis (specifically Hypothesis I (d) was a matching technique, in which judges were requested to blind-match two Rorschach records for each person. The use of this method has also been cited above in the review of previous studies (22), (50). The blind-matching procedure has been a favorite method for evaluating Rorschach data because it permits study of each record as a molar unit. Cronbach states that, "We can now obtain adequate evidence on the stability of Rorschach patterns only by such a method...". The liabilities of this method lie largely in the "human limitations of judges". As Cronbach points out, mismatching may occur because of "minor false elements. Matching, on the other hand, might be excellent, even perfect; the study would still not guarantee that each element...was correct, especially if the subjects were quite different from each other" (10).

The purpose of using this method in the present research was to take advantage of the more holistic approach available through judges' matchings and, more importantly, to demonstrate that the consistency provided by the response-comparison technique reflects significant aspects of the personality. It was felt that the deficiencies of this method was indicated by Cronbach might be resolved by careful design of the judge matching method.

It was assumed that mismatching because of "minor false elements" would be distributed by chance and hence should be equally distributed among the three groups tested, if as Hypothesis I (d) states, "Successful matchings...will be independent of the length of time between tests". In other words the emphasis here was on differences among three coefficients of contingency derived from judges' successes in matching pairs in the three groups rather than on one specific coefficient for the entire sample. On the other hand, care was taken to avoid having matchings made solely because of obvious differences among the subjects. This was done by submitting records of several patients in one set with the productivity (total number of responses) approximately equal for each patient in the set. Since productivity is one of the more obvious differences among subjects, e.g. it would be relatively simple to differentiate one subject with a response total of twenty from another subject with a

response total of sixty, it was felt that this means of control at least partially satisfied Cronbach's criticism.

The procedure followed in the judge-matching method was to submit the sixty pairs of records in twelve sets of five pairs each. Each set included at least one pair of records from each group. Each record contained only the responses, listed sequentially with their location specified according to Beck (5) and the position of the card specified for each response. The scoring of the responses was not included. Only the specific unelaborated content of the response was listed, thus eliminating any cues based on idiosyncratic verbalizations. Furthermore, as indicated above, each set of five pairs were within a specified range, based on the productivity. Four judges, each a staff member of the clinical psychology section of the Dearborn VA Hospital and with at least five years of Rorschach experience, performed the matchings.

At the conclusion of the procedure each judge submitted a verbal report of his impressions of factors involved in the matchings.

Questionnaire. A supplementary method used to obtain further information related to the effects of practice or memory on the retest was a list of ten questions (see Appendix A). This questionnaire constituted a structured verbal report of each subject's own perception of the effect of his recall on his responses in the retest. The questions

were not designed to provide quantitative data, but rather to give qualitative information for additional insight into the nature of the recall findings supplied by the recall technique.

D. Treatment of Data

In order to test the hypotheses of this investigation a series of measures were devised based on tabulations of the matching and recall data. Each measure will be discussed as it pertains to the particular hypothesis it was constructed to test. (See Table III for summary of measures).

The reporduction measure was designed to test Hypothesis I (a) and is a mean percentage derived by dividing each subject's total number of response reproductions by his total number of responses in the initial test; r/R(I). It represents the proportion of the total responses in a subject's initial test that he actually reproduced in the identical area in the retest. It may be considered a measure of consistency and is therefore hypothesized as being independent of the time variable.

The total reproduction measure was designed to test Hypothesis I (b) and is a mean percentage derived by dividing each subject's total number of reporductions, regardless of location, by his total number of responses in the initial test; r/rl/R(I). It represents the proportion of the total possible responses in a subject's initial test that he

actually reproduces in the retest but not necessarily in the identical location. It also may be considered a measure of consistency and hence unvarying over the time intervals specified.

The new response measure was designed to test Hypothesis I (c) and is a mean percentage derived by dividing each subject's total number of new responses in the retest by his total number of responses in the retest; nr/R(II). It represents the proportion of the total number of responses in a subject's retest that were not similar in any way with the initial test. It may be considered a measure of inconsistency but remaining constant over the varying short time intervals.

The reproduction-recall measure was designed to test Hypothesis II (a) and is a mean percentage derived by dividing each subject's correctly recalled reproductions, by his total number of response reproductions; cr/r. It represents the proportion of the responses of the initial test which were reproduced in the identical area of the retest and were correctly recalled by the subject as having been given in the initial test. It may be considered a measure of recall and was therefore hypothesized to decrease as a function of the time between tests.

The total reproduction-recall measure was designed to test Hypothesis II (b) and is a mean percentage derived by dividing each subject's correctly recalled reproductions

plus correctly recalled reproductions, location changed by his total number of response reproductions plus response reproductions, location changed; cr/crl/r/rl. It represents the proportion of the responses of the initial test which were reproduced, regardless of location, in the retest and were correctly recalled by the subject as having been given in the initial test. It is also a measure of recall and hence expected to decrease as a function of the length of time between tests.

The new response identification measure was designed to test Hypothesis II (c) and is a mean score derived by the sum of each subject's correctly identified new responses plus a constant of ten minus the sum of his questioned new responses plus incorrectly identified new responses; cnr/10-(qnr/inr). This measure varied from the mean percentages used for the measures described thus far because it was found that some subjects did not correctly identify any of their new responses. Hence the score was devised to take account of the varying totals of new responses which would not otherwise be indicated by zero percentage scores for these subjects. This score also measures recall and is similarly hypothesized to decrease as a function of the length of time between tests.

The recall measure was designed to test Hypothesis II (d) and is a mean percentage derived by dividing each subject's sum of correctly recalled reproductions plus

correctly recalled reproductions, location changed plus correctly identified new responses by his total number of responses in the retest; cr/crl/cnr/R(II). It represents the proportion of the total responses in a subject's retest that he actually correctly recalled as having been given in the initial test, regardless of location, or correctly identified as new responses appearing only in the retest. It is a more general measure of recall, being based on all of the responses of the retest, and is also predicted to decrease as a function of the length of time between tests.

Hypothesis I (d) was tested by the use of Vernon's formula[1] (52) which yielded three coefficients of contingency, one for each group. These coefficients were tested for significance of differences from one another. Confirmation of the null hypothesis with respect to these differences would be considered as another indication of the stability of the Rorschach over the short time intervals specified.

---

[1]See appendix C for Vernon's formula.

TABLE III

SUMMARY OF SCORING MEASURES

Related Hypothesis	Measure	Score	Definition
I(a)	Reproduction	$\frac{cr \neq qr \neq ir}{R(I)}$	$\frac{\text{all reproductions}}{\text{total responses in first test}}$
II(a)	Reproduction-recall	$\frac{cr}{cr \neq qr \neq ir}$	$\frac{\text{correctly recalled reproductions}}{\text{all reproductions}}$
I(b)	Total reproduction	$\frac{r \neq rl^{*}}{R(I)}$	$\frac{\text{all reproductions, regardless of location}}{\text{total responses in first test}}$
II(b)	Total reproduction-recall	$\frac{cr \neq crl}{r \neq rl}$	$\frac{\text{correctly recalled reproductions regardless of location}}{\text{all reproductions, regardless of location}}$
I(c)	New response	$\frac{cnr \neq qnr \neq inr}{R(II)}$	$\frac{\text{all new responses}}{\text{total responses in retest}}$
II(c)	New response-identification	$cnr \neq 10 - (qnr \neq inr)$	correctly identified new responses plus ten minus the sum of questioned new responses plus incorrectly identified new responses
II(d)	Recall	$\frac{cr \neq crl \neq cnr \neq crj}{R(II) \neq rj}$	$\frac{\text{correctly recalled reproductions, regardless of location, plus correctly identified new responses plus correctly identified rejections}}{\text{total responses in retest plus rejections in retest}}$

*$r \neq rl = (cr \neq qr \neq ir) \neq (crl \neq qrl \neq irl)$

IV. RESULTS

The statistical treatment of the seven measures derived from the response-comparison and recall data involved a comparison of group means for significant differences. The statistical technique employed for this comparison was the t-test. Since six out of the seven measures were based on percentage scores, a transformation was necessary to render the means and variances independent of one another (cf. Snedecor, p.446). The suitable transformation in this case is the inverse sine or angular transformation.[1]

Although the t-test applied to psychological data has come to be viewed critically by many psychologists in recent years, its use was felt to be appropriate for this study. The most general criticism of this technique is directed at the assumption of normality required for its use, an assumption which is questionable when applied to much of psychological data. Edwards (11) points out, however, that the two-tail t-test is relatively little influenced by departures from normality.

For convenience, the results pertaining to Hypotheses I (a) and II (a), I (b) and II (b) and I (c) and II (c) will be presented together since the measures, on which the tests of these hypotheses were based, are related.

---

[1] Although conclusions reached by analysis of original data are usually the same as that of transformed data, the test is made more sensitive in terms of the probability attached to the t-score obtained with transformed data (cf. Edwards, p.166)

The confirmation of Hypothesis I (a) requires that there be no significant differences between the mean percentages of response reproductions of any pair of the three groups. In other words the percentage of response reproductions should be constant with the passage of time. Table IV provides a comparison of groups I, II and III on differences in the reproduction measure. The findings

TABLE IV

COMPARISON OF GROUPS ON DIFFERENCES IN THE REPRODUCTION MEASURE

Groups	t	p
I - II	1.916	$<.10$
I - III	2.458	$<.02$
II - III	.730	$<.50$

indicated that the mean percentage (transformed) of response reproductions of group I tended to be different from that of group II ($P<.1$) and was significantly different from that of group III ($P<.02$). The means of groups II and III were not found significantly different ($P<.5$). These findings may be interpreted as indicating that the percentage of response reproductions tend to decrease after the short time interval of four hours but apparently becomes stabilized after two weeks for the period of time covered by the design of this research. The results thus partially confirm Hypothesis I (a).

The confirmation of Hypothesis II (a) requires that there be significant differences between the mean percentages of correctly recalled reproductions of any pair of the three groups. It further requires that the differences be positive when the mean of the group with the longer time interval between test and retest is subtracted from the group with the shorter time interval, e.g. mean of group I minus the mean of group II should bear a positive sign. In other words, the percentage of response reproductions which are correctly recalled should decrease with the passage of time. Table V provides a comparison of groups I, II and III on differences in the reproduction-recall measure. The findings indicated

TABLE V

COMPARISON OF GROUPS ON DIFFERENCES IN THE REPRODUCTION-RECALL MEASURE

Groups	t	p
I - II	3.137	$<.01$
I - III	5.821	$<.001$
II - III	2.475	$<.02$

that the mean percentage (transformed) of correctly recalled reproductions of group I was significantly different from that of group II ($P<.01$) and from that of group III ($P<.001$). The differences between the means of groups II and III were also significantly different ($P<.02$). The differences were all in the predicted direction. It seems to be a safe assumption that recall of response reproductions decreases

as a function of time. The results thus confirm Hypothesis II (a).

When the findings of the two measures reported above are compared, one may observe that while response reproductions tended to remain constant from two weeks up to two months, the recall of those response reproductions continued to decrease.

The confirmation of Hypothesis I (b) requires that there be no significant differences between the mean percentages of all response reproductions, regardless of location, of any pair of the three groups. In other words, the percentage of all repeated responses should not vary with the passage of time. Table VI provides a comparison of groups I, II and III on differences in the total reproduction measure. The findings indicated that the mean percentage

TABLE VI

COMPARISON OF GROUPS ON DIFFERENCES IN THE TOTAL REPRODUCTION MEASURE

Groups	t	p
I - II	1.518	$<.20$
I - III	1.594	$<.20$
II - III	- .076	$>.90$

(transformed) of response reproductions plus response reproductions, location changed of group I was not significantly different from that of group II ($P<.2$) or group III ($P<.2$), nor were the means of groups II and III significantly different ($P>.9$). Despite the lack of statistical significance between the means of groups I and

II, the differences were in the direction of decreasing mean percentages as time increased, the same direction as in the case of the response reproductions reported above. (There is, of course, some overlap between the two measures since both involve response reproductions). Thus, these results are also suggestive of a decreasing tendency after the four hour test-retest interval and show more stability from the two week retest interval up to two months. However on a statistical basis the results may be considered as confirmatory of Hypothesis I (b).

The confirmation of Hypothesis II (b) requires that there be significant differences, between the mean percentages of the sum of correctly recalled reproductions plus correctly recalled reproductions, location changed of any pair of the three groups. As in the case of Hypothesis II (a) direction is also predicted requiring that the signs all be positive. In other words, the percentage of all repeated responses, regardless of location, which are correctly recalled should decrease with the passage of time. Table VII provides a comparison of groups I, II and III on differences in the total reproduction recall measure. The

TABLE VII

COMPARISON OF GROUPS ON DIFFERENCES IN THE TOTAL REPRODUCTION-RECALL MEASURE

Groups	t	p
I - II	2.121	$<.01$
I - III	5.504	$<.001$
II - III	2.733	$<.01$

findings indicated that the mean percentage (transformed) of correctly recalled reproductions, less correctly recalled reproductions, location changed of group I was significantly different from that of group II ($P<.01$) and from that of group III ($P<.001$). The means of groups II and III were also significantly different ($P<.01$). Differences were all in the predicted direction. The findings thus indicate that recall of all reproduced responses decreases as a function of time and confirm Hypothesis II (b).

When the findings of the two latter measures reported above are compared to those of the former, the results appear quite similar. Consistency is even more pronounced in total reproductions, with the difference in means between the four hour and two month groups falling short of significance at the 20% level. In both measures of consistency from four hours to two weeks, however, there is a tendency to vary in the same direction as the measures of recall. Both measures of recall show a definite decline in recall as a function of time.

The confirmation of Hypothesis I (c) requires that there be no significant differences between the mean percentages of new responses of any pair of the three groups. In other words the percentage of new responses produced on retest should not vary as a function of time. Table VIII provides a comparison of groups I, II and III on differences in the new response measure. The findings indicated that

TABLE VIII

COMPARISON OF GROUPS ON DIFFERENCES IN THE NEW RESPONSE MEASURE

Groups	t	p
I - II	-1.630	$<.20$
I - III	-1.378	$<.20$
II - III	.506	$<.70$

the mean percentage (transformed) of new responses of group II ($P<2$) or group III ($P<2$), nor were the means of groups II and III significantly different ($P<.7$). The lack of significant differences between the means of the three groups suggests that the amount of change remains constant with the increase in the elapse of time between test and retest from four hours up to two months. The results confirm Hypothesis I (c).

The confirmation of Hypothesis II (c) requires that there be significant differences between the mean scores of correctly identified new responses of any pair of the three groups. As in the case of the other two measures of recall, direction is also predicted requiring that the signs all be positive. In other words the number of new responses which are correctly identified as new responses should decrease with the passage of time. Table IX provides a comparison of groups I, II and III on differences in the new response-identification measure. The findings indicated that the mean score of correctly identified new responses of group I

TABLE IX

COMPARISON OF GROUPS ON DIFFERENCES IN THE NEW RESPONSE-IDENTIFICATION MEASURE

Groups	t	p
I - II	1.534	$<.10$
I - III	5.760	$<.001$
II - III	2.171	$<.05$

was different from that of group II at the ten percent level and was significantly different from that of group III ($P<.001$). The means of groups II and III were significantly different at the 5% level. The differences were all in the predicted direction. It is not clear why the t-scores of this measure were reduced from those of the other two measures of recall. The results partially confirm Hypothesis II (c).

A comparison of the findings of the two measures based on new responses with the findings of the other measures reveals basically similar results. The new response measure as an indicator of inconsistency is the logical complement of the measures of consistency and should be equally consistent. The new response-identification measure as a measure of recall produced generally similar results to those of the other measures of recall.

The confirmation of Hypothesis II (d) requires that there be significant differences between the mean percentages of the sum of correctly recalled reproductions, regardless

of location plus correctly identified new responses of any pair of the three groups. The prediction of direction requires that the signs all be positive. In other words, this is a general measure of recall and should yield results indicating decrease with the passage of time. Table X provides a comparison of groups I, II and III on differences

TABLE X

COMPARISON OF GROUPS ON DIFFERENCES IN THE RECALL MEASURE

Groups	t	p
I - II	4.504	$<.001$
I - III	7.366	$<.001$
II - III	2.990	$<.01$

in the recall measures. The findings indicated that the mean percentage (transformed) of the sum of all correctly recalled reproductions, regardless of location, plus correctly identified new responses of group I was significantly different from those of group II ($P<.001$) and group III ($P<.001$). The means of groups II and III were also significantly different ($P<.01$). The differences were all in the predicted direction. The findings thus indicate that recall in the Rorschach decreases as a function of time and confirm Hypothesis II (d).

The foregoing have been statistical tests of the hypotheses based on the response-matching and recall data.

The statistical treatment of the group-matching data involved a t-test of differences between coefficients of contingency. The error term employed is also given by Vernon's formula (see appendix C). The statistical test of Hypothesis I (d) requires that there be no significant differences between the coefficients of contingency for successful matchings of pairs of response records of any pair of the three groups. In other words successful matching by judges of two tests from the same individual should be independent of the length of time between tests. The results were perfect matchings by each judge of all sixty pairs of records. All of the judges pointed out that idiosyncratic content was largely the basis for their matchings with occasional reliance on patterns of responses. Even in the latter instance pairing was done on the basis of a few cards, not necessarily the complete record. In summary, the judge matching method as carried out in this study is still subject to Cronbach's criticisms (see page 40) despite the precautions taken to avoid them. Hence the results pertaining to Hypothesis I (d) may be considered inconclusive.

Table XI presents the mean percentages derived from each of the seven measures for the three groups. Several of these percentages are of particular interest. The mean percentages of response reproductions for the three groups

TABLE XI

COMPARISON OF GROUPS ON MEAN PERCENTAGES OF ALL MEASURES

Measure	I	II	III
Reproduction	72.3%	63.4%	60.2%
Reproduction-Recall	99.2%	94.8%	88.1%
Total Reproduction	76.1%	69.4%	70.8%
Total Reproduction-Recall	95.1%	87.2%	75.3%
New Response	25.0%	32.2%	29.5%
New Response-Identification	75.5%	52.8%	32.3%
Recall	91.0%	76.5%	62.0%

are 72.3, 63.4 and 60.2 respectively. This might be compared to the results of Swift[1], who found 57% identical responses in a retest after two weeks and 51% after thirty days among pre-school children (48). Allen, Hamm and Stiff[2] found mean percentages after six weeks in a range from 27 to 34.6 of "consistent" responses in a retest series of both standard and achromatic cards (2). The related statistics in the present study are all above those reported in the literature.

Other findings presented on this table, which are of interest, are the mean percentages of new responses in the retest. For the three groups, these percentages were 25, 32.2 and 29.5 respectively. In other words, after four hours 25% of the responses in the retest are different from any of those in the initial test and after two months the same percentage is 29.5. (The difference has been indicated in Table VIII as being non-significant.)

Tables XII through XIX (see appendix B) present summaries of subjects' responses to the questionnaire. These data will be discussed in the following section.

---

[1]See page 12 of this thesis.
[2]See page 21 of this thesis.

V. DISCUSSION

A. Consistency and Recall

Controversy has long clouded the effects of recall in retesting with the Rorschach. Whereas Thornton and Guilford (47) emphasize that retesting must be avoided because of memory, Piotrowski (37) claims that practice effects are unimportant because there is no conscious effort and that repeated responses represent not recall but rather evidence of the stability of personality for the period of time intervening between tests. Hons (34) states that "at ... times a person may believe he has produced a new response entirely, yet a comparison with the earlier record shows that it was merely forgotten".

It seemed feasible to test the argument that consistency between test and retest responses is not solely accounted for in terms of memory. Accordingly two hypotheses were proposed: 1) Performance on the Rorschach in terms of repeated responses on retest remains constant over varied short time intervals; 2) The effect of memory of the first test on the retest in terms of those responses would decrease as a function of the length of time between tests.

The results on the whole confirmed these hypotheses. The measures of consistency devised to test the first hypothesis, with one exception, did not show significant

differences among groups, retested at varied lengths of time. The measures of recall devised to test the second hypothesis decreased as a function of the length of time between tests, but also with one exception. These findings would tend to support the contention that retest consistency is not to be solely accounted for in terms of recall.

It should be noted that the results also indicated a tendency (not statistically significant) of the measures of consistency to vary in the same direction as those of recall, i.e. decreases with the passage of time. This tendency was apparent from the four hour to the two week retesting interval but not thereafter. This might suggest that the consistency in retest findings may be related to retention over test-retest intervals ranging up to two weeks but not thereafter.

B. Some Conditions Related to Consistency

It was pointed out in the previous section that the percentages of reproduced responses in the present study were all above those reported in the literature. The testing conditions for this study all tended to maximize these consistencies. In this respect the following conditions are those that might be considered as contributory toward maximal consistency: 1) all testing by same examiner; 2) the constant environment represented by hospitalization;

3) the exclusion of neuropsychiatric patients; 4) no exposure to any psychological or psychiatric treatment between tests; 5) the use of adults as subjects; 6) the use of standardized instructions; 7) the short test-retest time intervals.

Some indication of recall as a condition in the consistency of responses may be derived from the verbal reports of the subjects. Table XIV, which indicates opinions on whether deliberate efforts were made to repeat responses, reveals that 40% or 24 of the subjects report that it was the stimulus properties of the cards that tended to elicit the same responses. Characteristic statements of this type expressed the belief that responses were repeated because "they were there" or "saw them whether I thought about it or not". This may be compared to 13% or eight of the subjects who reported that recall elicited the same responses on retest. Thus the subjects' own perception of their recall seems to support the claim that consistency of response is due to more than retention.

C. Some Conditions Related to Inconsistency

One of the purposes of this study was to evaluate the changes in the responses which are a function of the passage of time. Table XI presents findings related to this. It reveals what seems to be a surprisingly high number of new responses (mean percentage of 25) on retest after four hours.

After two months this percentage is only 29.5. Here also the verbal reports of the subjects are helpful in determining the conditions related to inconsistency.

Table XVI gives information related to inconsistency as represented by the deliberate attempt to produce new responses. Categories A and B combined reveal that 15 of group I subjects as compared to 14 of group II and eight of group III subjects reported making an effort to give new responses out of curiosity or in order to be more thorough.

Table XVIII, which indicates subjects' points of view on the effects of recall, reveals on the whole, a perception of the retest as a more relaxed situation, where practice and familiarity changed the manner of responding. Category B indicates that 14 subjects in group I as compared to six subjects in group II and two in group III reported that having taken the initial test made it easier to concentrate and look for more percepts in the retest. The subjects also stressed the effects of recall in the retest and the reduction in the ambiguity of the test situation, both of which tend to produce consistency. Thus, the altered situation of the retest gives rise to both inconsistency and consistency with the initial test.

Tables XV and XVII reveal that for a variety of reasons subjects deliberately suppress responses. Whatever the reason, the deliberate suppression of responses contributes

to inconsistency. The preponderance of subjects who reported doing this appear in group I.

In conclusion many of the conditions which seem to account for inconsistency appear to be most prominent in the four hour retest group. It is suggested that this may explain the relatively high degree of inconsistency after four hours where one would expect much less variability.

D. Judges' Verbal Reports

Despite the fact that the results of the judge matching procedure precluded any statistical treatment, the verbal reports of the judges are useful in shedding additional light on the nature of these findings. Although all of the judges reported a subjective impression of general reliability in the response patterns, they also indicated that there were a number of records that could be matched only by elimination. In other words, the reliability of certain response patterns appeared reduced from those found in the majority of the records.

It is suggested that these especially inconsistent records might be due to an extreme combination of the suppression of responses as mentioned previously, varying degrees of motivation, misinterpretations of instructions, degree of threat by the task or of the many other reasons listed in Tables XIII, XVI and XVIII. Such idiosyncratic reactions to testing emphasize the importance of attempting

to gain insight into what Schachtel (45) calls the "subjective definition" given by the subject to the testing situation. He points out that every subject defines the test situation in terms of his own needs, wishes and fears. He suggests that one might deal with the "subjective definition" either by attempting to minimize the influences which give rise to it or by evaluating it as an additional source of insight into the subject's personality and attitudes.

E. A Theoretical Interpretation

It should be mentioned that the findings of the present study do not necessarily support the contention of Piotrowski that consistency is due to the stability of the personality in the time intervening between tests. These findings simply suggest that recall is not the sole basis for consistency. Further studies of validity would be necessary to substantiate his claim.

On the other hand, an alternate approach to the results of this study might be pointed out. It may be said that in the Rorschach as in any test which is purported to tap personality, the findings of both consistency and inconsistency are anticipated by personality theory. The organization of personality is not static and hence undergoes shifts and changes of emphases even within its relatively stable framework. Hence, even fluctuations within the day, as represented by changes in mood, attitudes, etc., are to be found along with the continuous consistent core of organization.

F. Methodology

The methodology employed in this investigation might warrant some attention. The choice of subjects for a sample raises the question of the larger population represented by the sample to which the findings can be generalized. The choice here was a practical one and may be said to be representative of an extensive hospitalized veteran population which is the focus of the professional activity of a considerable number of clinical psychologists. It is doubtful whether a clinical instrument such as the Rorschach can ever be adequately tested as a clinical instrument in a sample designed to be representative of the general population. The difficulty lies in presenting the test as it normally would be in a clinical situation. As indicated above the "subjective definition" which is of much importance would be quite different if the test were presented in a "research" as opposed to a clinical setting. In the present research the test was presented to each subject in a clinical setting which was in no way different from that of a routine referral for diagnostic testing.

G. Implications for Further Research

A number of implications for further research are suggested by this study. The most obvious would be to extend the range of the temporal dimension so as to include much longer test-retest intervals. It would be valuable to determine where the retention curve in terms of the recall

measures reaches a plateau and also the point in time where the consistency measure decreases. Another suggestion for future investigation would be the application of the various measures employed in this study to the determinants. One might ask, for example, the extent to which recall is represented in the consistency of the use of color, form or combinations thereof. A third suggestion would be the personality correlates of changes in the retest as indicated by the new response measure. These changes may be related to the "stability score" devised by Gibby (16) by means of which he was able to determine those individuals in experimental retest groups, who were able to effect most change under experimentally altered conditions. The changes indicated in this study by the new response measure were of a spontaneous nature, stemming from subjects' own interpretation of the retest situation. Research interest would also be attached to the type of responses that persist, change or drop out and the personality correlates thereof.

VI. SUMMARY

A review of the literature on Rorschach test-retest studies indicated that research has never been focused on temporal reliability, i.e. the consistency of retest findings over varying periods of time. Furthermore the test-retest approach to reliability has been considered questionable because of practice or memory effects. Rorschach writers have disagreed on the importance of recall effects. Some argue that retesting is meaningless because of memory whereas others insist that repeated responses are representative of stable personality trends rather than of recall.

The purpose of this study was to investigate the consistency of Rorschach results where subjects were not exposed to any treatment other than the systematic varying of time intervening between tests. Specifically, it was to determine the differential effects of the passage of time on the persistence, change and recall of Rorschach responses on retest.

Two hypotheses were set forth: 1) performance on the Rorschach in terms of repeated responses on retest remains constant over varied short time intervals; 2) recall of those responses decreases as a function of the length of time between tests.

Sixty patients screened to assure exclusion of those with neuropsychiatric conditions were selected as subjects from a VA general medical and surgical hospital. The subjects were distributed into three groups of twenty, equated for age and intelligence. Each group was retested with the Rorschach after the following approximate time intervals: group I, four hours; group II, two weeks, and group III, two months. Following the retest of each subject, his responses were individually read back to him and he was asked if these responses had been given in the initial test. Following the recall procedure a questionnaire, designed to obtain a verbal report of the effect of recall on the retest, was administered to each subject.

Two techniques were used to obtain data to test the hypotheses. The response-comparison technique was a matching procedure, by means of which each pair of Rorschach protocols were compared for common or consistent responses. The recall technique was a scoring method, by means of which accuracy in the identification of retest responses as new or repeated responses was determined. The two techniques yielded seven measures which were tested for significance of differences among the three groups by the t-test.

The results on the whole confirmed both hypotheses. It was found that the measures of consistency devised to test the first hypothesis did not yield significant differences

among the three groups, regardless of the length of time elapsed before retest. The measures of recall devised to test the second hypothesis decreased as a function of the length of time between tests. It is concluded on the basis of the results that retest consistency is not to be solely accounted for in terms of recall.

It was also noted that measures of consistency tended to decrease in the same direction of the measures of recall for the time intervals from four hours to two weeks.

The findings in terms of the percentages of consistent and new responses on retest were presented and discussed. It was pointed out that the percentages of response consistency were above those reported in the literature. A number of conditions were suggested as related to the enhanced test reliability indicated by the present findings. Also, some conditions related to inconsistency as suggested by the verbal reports were discussed.

The results were discussed from the point of view of their relevance to Rorschach reliability and to personality theory.

The choice of subjects used in this study also received attention.

Finally, implications for further research, stemming from the present study, were reviewed.

APPENDIX A

QUESTIONNAIRE

1. Did you expect to take this test again?

2. Why do you think this test was given again?

3. Did you make an effort to see the same things this time that you saw the first time that you took the test? Why?

4. Did you avoid mentioning anything this time that you saw the first time? Why?

5. Did you make an effort to see anything new this time? Why?

6. Did you see and mention anything this time that you saw the first time that you took the test but did not mention at that time? What and why?

7. Did you mention anything the first time that you did not mention this time? Why?

8. Did you mention anything this time that you did not mention the first time? What?

9. Did you see more things the first time or this time?

10. What do you feel are the effects of taking the first test on this test? What do you think the effects of taking the first test would be on the second test if it were given ten years later instead of at this time?

APPENDIX B

TABLE XII

COMPARISON OF GROUPS ON REPLIES TO QUESTIONS 1,3,4,5,6,7,8,9

No.	Question	Yes			No			Don't know		
		I	II	III	I	II	III	I	II	III
1.	Expect to take test again?	1	0	1	19	20	19	0	0	0
3.	Make effort to see same things?	6	8	9	14	12	10	0	0	1
4.	Avoid mentioning anything seen first time?	8	4	2	11	16	18	1	0	0
5.	Make effort to see anything new?	18	20	18	2	0	2	0	0	0
6.	Mention anything this time, seen but not mentioned first time?	4	3	1	16	15	18	0	2	1
7.	Mention anything first time, not mentioned this time?	14	6	6	4	10	5	2	4	9
8.	Mention anything this time, not mentioned first time?	19	17	16	1	1	4	0	2	0
9.	See more things the second time?	18	13	7	0	1	6	2*	6*	7*

*Subjects replied that the number of responses was the same in both tests.

TABLE XIII

COMPARISON OF GROUPS ON QUESTION #2

(Why test was given again?)

Categories of Reasons	Totals * I	II	III
A. Experimental	2	3	1
see if imagination depends on memory	1		
determine effect of medicine	1		
see if there is the same reaction to the test twice		1	
see if one is struck by same initial impression		1	
determine different frames of mind		1	
determine effects of hospitalization			1
B. Clinical	8	4	0
check sanity	1		
check for lying	1		
thinking ability and concentration	1		
check mental functioning	3	1	
indication of instability	1	1	
test observance at first sight, paying attention	1	1	
checkup		1	
C. Memory	18	10	8
determine if one sees same things	7	1	4
check memory	11	9	3
learning ability			1
D. Changes	16	10	10
comparison	1	8	4
determine changes, differences, added things; changes of opinion	15	2	6

TABLE XIII (continued)

Categories of Reasons	Totals I	II	* III
E. Part of test procedure	4	3	4
broke the ice (with the first test)	1		
more conclusive	1	1	
more certain of your information	1		
obtain pattern (with both tests)	1		
check accuracy of first test		1	
see how good my first impression is		1	
confirm first findings			1
routine			1
part of original test			1
clarification of first test			1
F. Other or no opinion	2	5	8
no opinion	2	4	6
might have fouled up last time		1	
something went wrong			1
test my sincerity			1

*Some subjects gave more than one reason which accounts for sums of totals equaling more than total N. of twenty for each group.

TABLE XIV

COMPARISON OF GROUPS ON QUESTION #3

(Make effort to see same things?)

	Categories of Reasons	Totals I	II	III
A.	No effort because of stimulus properties of cards	9	8	7

"can't help seeing same things"
"saw them whether I thought about it or not"
"they were there, no effort, would see same things without memory"
"just what came to me, tried to find something else, but it still looked the same, because it's there"
"just saw same things without thinking about it, but it came to my mind that I'd seen it last time"

B.	No effort because of memory	2	3	3

"wanted to see how much I remembered, could have made it look like something else"
"no special effort, but hard not to say them when you remember them"
"same things came back easily, one would remind me of another and I'd go looking for it and see other things in the process too"
"had in my mind what I saw first time and naturally saw it right away again the second time"

C.	Merely carried out instructions to report everything seen	4	3	5

"just said whatever came to my mind"
"just tried to see everything I could see, made no special effort"
"did recall but did not let it lead me to see what I did this time"
"I didn't purposefully attempt to, may have unconsciously"

TABLE XIV (continued)

Categories of Reasons	Totals I	II	III
D. Made an effort to see same things	1	2	2
"thought it was a checkup and I wanted to be as near right as possible"			
"realized it was the same test; should see same things, maybe more realistically"			
"wanted to know if I could and let you know; looked for same things again"			
E. Looked for something different and other reasons	4	4	3
"made no attempt; wondered if I could see something different"			
"tried to look for more things; things don't look the same every time you look at them"			
"tried to see more for self satisfaction"			
"you asked me to see same things"			

TABLE XV

EXPLANATIONS OF "YES" TO QUESTION #4 OF QUESTIONNAIRE

	Reasons for "yes" answers	Total
Group I	"didn't think the second time that it was what I said the first time" "'skeleton of bat' didn't take the shape the second time" "if that was a 'man' I'd need glasses" "some things looked different: I saw them but they didn't look like it, so I didn't mention them"	4
Group II	"skeptical of one answer, too vague" "some didn't make sense this time" "some minor points of elaboration such as sexual organs; you didn't ask about them" "one thing didn't strike me as such"	4
Group III	"mentioned 'woman without a head' first time, seemed too silly"	1

TABLE XVI

COMPARISON OF GROUPS ON QUESTION #5

("Yes" - Attempted to give new responses)

Categories of Reasons	Totals I	II	III
A. Might have overlooked something; more thorough	8	6	4

"wanted to see if I overlooked something"
"felt I wasn't seeing enough, tried hard to concentrate"
"didn't want to miss any hidden object"
"I thought about last time, didn't see things I should have"

B. Curiosity; looked for something different or more	7	8	4

"wanted to find something different, find what inkblots represented"
"felt someone else would see something different"
"mostly curiosity, to see if there was anything else"
"thought it would be interesting to see something else"
"tried to see if I was more alert, tested to see how much could be gotten out of it"

C. Complied with instructions to see everything	2	4	4

"wanted to make sure I saw everything"
"tried to see everything I could, as you said"
"made an effort to see exactly what was on there"
"tried to see all I could, didn't feel too good last time"

D. Instructed to see more	2	1	2

"thought you asked me to see if there was anything I didn't see the first time"
"see if missed anything last time, you said to do it"
"felt that's what I was supposed to do"

TABLE XVI (continued)

Categories of Reasons	Totals I	II	III
E. Didn't try or unable to see more	1	0	3
"seemed to be same things" "couldn't imagine anything new" "made an effort, but don't think it was successful"			
F. Other reasons	0	1	3
"yes, because test was given to see if I changed my opinions" "wanted to add to my intelligence or imagination, it might have some bearing on my intelligence" "wanted to increase artistry" "thought I should change my mind after two months"			

TABLE XVII

REPLIES OF "YES" TO QUESTION #6 OF QUESTIONNAIRE

(Gave responses on retest that were seen but not mentioned in initial test)

	Reasons for "yes" answers	Total
Group I	"figured I had seen enough" "thought (a response) was silly" "thought (a response) was of no importance" "didn't see (a response) until after I gave back the card"	4
Group II	"(response) wasn't as apparent the first time"	1
Group III	none	0

TABLE XVIII

COMPARISON OF GROUPS ON QUESTION #10(a) OF QUESTIONNAIRE

(Effects of first test on retest)

Categories of Replies	Totals I	II	III
A. Familiarity; knew what to expect; could respond faster	8	7	5
"didn't take as long to grasp what was on the cards" "more familiar, more at ease, better idea of what to do" "more nervous first time, knew more about it this time" "nervous and afraid to say some things the first time, because it might be wrong or not really there"			
B. Ease of concentration; could look for more things	14	6	2
"helped a lot, concentrated on what I'd missed" "don't have to look too hard second time" "saw previous things first, went into a little more detail" "easier, knew some things already there, so looked for new ones" "paid more attention to outline and details"			
C. Tended to see same things due to recall effects	7	8	5
"what you see before stays in your mind" "made me look for things I saw last time rather than look fornew ones" "automatically see again same things, fresh in my mind from last time" "popped into your mind that you'd see it earlier"			

TABLE XVIII (continued)

Categories of Replies	Totals I	II	III
D. No effect; recall effects negligible	0	8	9

"looked at cards as if for first time"
"would have seen same things even if I had not taken test previously"
"things popped out and I'm not sure whether I remembered them or saw them again"
"very little difference, saw about same things automatically"
"memory actually secondary since things looked the same, I seemed to be drawn to same things in same order"

E. Other	1	2	1

"didn't turn cards as much"
"easier first time because subject was fresh"
"forgotten all about it, put me to thinking about it"
"didn't have to look for as many things

TABLE XIX

COMPARISON OF GROUPS ON QUESTION #10(b) OF QUESTIONNAIRE

(Possible effects of first test on retest 10 years later)

Categories of Replies	Totals I	II	III
A. Differences because of new experiences	11	8	6

"a lot would happen in ten years, change outlook; would imagine different things depending on intervening experiences"
"probably see different things: things always look different in time, because of nature probably would take different forms"
"could be a lot of things that would remind you of something different, you would associate them with things that will happen in the years from now"
"big difference; I'd be older, seen more things, the things I saw were related to things I've seen or experienced; in ten years I'd see more and different things"
"probably see different things; naturally when one sees an undescriptive picture he describes things from what he's familiar with; in ten years I'll be older and in a different situation or environment"

B. Differences because of no recall	3	3	2

"probably not remember, it would be like taking the test for the first time"
"great deal of difference, doubt if I'd remember cards at all"
"entirely different, like two different tests, would have forgotten what I saw ten years before, would be like new material"

TABLE XIX (continued)

Categories of replies	Totals I	II	III
C. Same things because of identical stimuli	2	3	3

"resemble same things from the shapes, which were all I had to go by, they would be the same in ten years"
"don't think it would make any difference, cards wouldn't change; I'd see some things in ten years if in same mental condition"
"no changes unless cards fade, depends on health and vision"
"I'll see same things because the shapes are outstanding and would convey the same meaning"

D. Same things because of recall	1	4	8

"not too different, I'll still remember it, I can remember well"
"no differences, ten years is not too long, can certainly remember in ten years"
"pretty close to the same things; it's always been that way for me, I get an impression and don't easily forget it"
"pretty close to same results; once you have learned something, it settles in your mind"

E. Same things regardless of recall	5	5	5

"might still see same things, but might not remember taking test before"
"be no retention at all; things you see are connected with your type of mind and mental health, if that didn't change, the things you'd see wouldn't change either"
"might see same things, but wouldn't go looking for them"
"no differences; probably forget first test and it would be like taking it first time; differences depend on frame of mind"

TABLE XIX (continued)

Categories of replies	Totals I	II	III
F. Other replies	1	0	2

"have same idea as in first test, trying to figure what is there instead of what actually is there; instead of looking. I'd try to recall first test and look for that"
"no idea"
"would see something different because the objects are not shaped true to form"

APPENDIX C

VERNON'S FORMULAE

FOR EVALUATING JUDGE'S MATCHINGS (52)

Coefficient of Contingency

$$C = \sqrt{\frac{(St-1)^2}{(t-1)\not(St-1)^2}}$$

Error Term

$$E = (t-1)\sqrt{\frac{(St-1)}{t}\left[\frac{(t-1)^2}{(t-1)\not(St-1)^2}\not 1\right]\not t\left[\frac{(t-1)-(St-1)^2}{3}\right]}$$

$$\sqrt{N}$$

t - number of elements to be matched

N - total number of judgements or matches

S - proportion of the judgements that are correct

BIBLIOGRAPHY

1. Ainsworth, M. D., "Problems of Validation" in Klopfer, B., Ainsworth, M. D., Klopfer, W. G., and Holt, R. R., Developments in the Rorschach Technique, Vol. I, World Book Company, N. Y., 1954

2. Allen, R. M., Manne, S. H., and Stiff, M., The influence of color on the consistency of response in the Rorschach Test, J. Clin. Psychol., 1952, 8, 97-98

3. Baughman, E., Rorschach scores as a function of examiner differences, J. Proj. Tech., 1951, 15, 243-249

4. Beck, S. J., Stability of the personality structure, Psych. Bull., 1942, 39, 512 (abstract)

5. _____, Rorschach's Test, Vol. I, Basic Processes, N. Y., Grune and Stratton, 1944

6. Bradway, K., Rorschach records of a schizophrenic patient before, during and after electric shock and insulin treatment, J. Proj. Tech., 1951, 15, 87-97

7. Brosin, H. W., and Fromm, E. G., Some principles of Gestalt Psychology in the Rorschach experiment, Ror. Res. Exch., 1942, 6, 1-15

8. Carp, A. L., and Shevzin, A. R., The susceptibility to falsification of the Rorschach psychodiagnostic technique, J. Consult., 1950, 14, 230-233

9. Counts, R. M., and Marsh, J. N., Personality characteristics in hypnotically induces hostility, J. Clin. Psychol., 1950, 6, 325-330

10. Cronbach, L. J., Statistical methods applied to Rorschach scores, a review, Psych. Bull., 1949, 46, 393-429

11. Edwards, A. L., Experimental Design in Psychological Research, Rinehart & Co., N. Y., 1950

12. Eicher, R. M., A comparison of the Rorschach and Behn-Rorschach Inkblot tests, J. Consult. Psychol., 1951, 15, 185-189

13. Ford, M., The Application of the Rorschach Test to Young Children, Univ. of Minn. Press, Minn., 1946

14. Fosberg, J. A., Rorschach reactions under varied instructions, Ror. Res. Exch., 1939, 3, 12-38

15. _____, An experimental study of the reliability of the Rorschach psychodiagnostic technique, Ror. Res. Exch., 1941, 5, 72-84

16. Gibby, Robert G., The stability of Certain Rorschach Variables under Conditions of Experimentally Induced Sets: I. The Intellectual Variables. J. Proj. Tech., 1951, 15, 3-26

17. Graham, Virginia T., Psychological studies of hypoglycemia therapy, J. Psychol., 1940, 10, 327-358

18. Griffith, Richard M., Test-retest similarity of the Rorschachs of patients without retention, Korsakoff, J. Proj. Tech., 1951, 15, 516-525

19. Halpern, F., Rorschach interpretation of the personality structure in schizophrenics who benefit from insulin therapy, Psychiat. Quarterly, 1940, 14, 826-833

20. Hertz, M., Reliability of the Rorschach Inkblot test, J. Applied, 1934, 18, 461-477

21. _____, Current problems in Rorschach theory and technique, J. Proj. Tech., 1951, 15, 307-338

22. Holzberg, J. D., and Wexler, M., The predictability of schizophrenic performance on the Rorschach test, J. Consult Psychol., 1950, 14, 395-399

23. Hutt, M. L., Gibby, R. G., Milton, E., and Pottharst, K., The effect of varied sets on Rorschach test performance, J. Proj. Tech., 1950, 14, 181-186

24. Jackson, R. W. B., and Ferguson, G. A., Studies on the Reliability of Tests, Univ. of Toronto Press, Toronto, 1951

25. Kelley, D. M. and Levine, K., Rorschach studies during sodium amytal narcoses (abstract), Ror. Res. Exch., 1940, 4, 146

26. _____, Margulies, H., and Barrera, S. E., The stability of the Rorschach method as demonstrated in electric convulsive cases, Ror. Res. Exch., 1941, 5, 35-43

27. Kerr, M., Temperamental differences in twins, Brit. J. Psychol., 1936, 27, 91-99

28. Kimble, G. A., Social influence on Rorschach records, J. Ab. & Soc. Psych., 1945, 40, 89-93

29. Kisker, G. W., A projective approach to personality patterns during insulin shock and metrazol convulsive therapy, J. Ab. & Soc. Psychol., 1942, 37, 120-124

30. Krout, J., Krout, M., and Dubin, T., Rorschach test-retest as a guage of progress in psychotherapy, J. Clin. Psych., 1952, 8, 380-384.

31. Lane, B., A validation test of the Rorschach Movement interpretation, Amer. J. Ortho., 1948, 18, 292-296.

32. Levine, K., Grasni, J. R., Gerson, H., Hypnotically induced mood changes on the verbal and graphic Rorschach, Ror. Res. Exch., 1943, 7, 130-134

33. Lord, E., Experimentally induced variations in Rorschach performance, Psych. Monog., 1950, 64

34. Mira, E., Sobre el valor del psicodiagnostico de Rorschach, Progressos de la Clinica, 1935, 30, 808-845

35. Hons, W. E., Principles and Practice of the Rorschach Personality Test, Lippincott, 1948

36. Norman, R. D., Shephard, L., Radlo, R., The influence of a superficial immediately preceding "set" upon responses to the Rorschach, J. Consult. Psych., 1952, 16, 261-264

37. Piotrowski, Z. A., Rorschach manifestations of improvement in insulin treated schizophrenics, Psychosomat. Med., 1939, 1, 508-526

38. _____, The reliability of Rorschach's Erlebnistypus, J. Ab. & Soc. Psych., 1937, 32, 439-445.

39. Rabin, A. I., "Validating and Experimental Studies with the Rorschach-Method," in Anderson, H., and Anderson, G., (eds.) An Introduction to Projective Techniques, N. Y., Prentice-Hall, 1951

40. _____, Nelson, W., and Clark, M., Rorschach content as a function of perceptual experience and sex of the examiner, J. Clin. Psych., 1954, 10, 188-190

41. _____, and Sanderson, M.H., An experimental inquiry into some Rorschach procedures, J. Clin. Psych. 1946, 3, 216-225

42. Rioch, M. J., The use of the Rorschach test in the assessment of change in patients under psychotherapy, Psychiat., 1949, 427-434

43. Rorschach, H., Psychodiagnostics (2nd Ed.), N. Y., Grune and Stratton, 1942

44. Sarbin, T. R., Rorschach patterns under hypnosis, Amer. J. Ortho., 1939, 9, 315-319

45. Schachtel, E. G., Subjective definitions of the Rorschach test situation and their effect on test performance. Contributions to an understanding of Rorschach's test III, Psychiatry, 1945, 8, 417-448

46. Singer, J. L., The Behn-Rorschach inkblots: A preliminary comparison with the original Rorschach series, J. Proj. Tech., 1952, 16, 238-245

47. Snedecor, G. W., Statistical Methods, ed. 4, Iowa State College Press, Ames, Iowa, 1946

48. Swift, J. W., Reliabilities of Rorschach scoring categories with pre-school children, Child Develop., 1944, 15, 207-216

49. Thornton, G. R., and Guilford, J. P., The reliability and meaning of Erlebnistypus scores in the Rorschach test, J. Ab. & Soc. Psych., 1936, 31, 324-330

50. Troup, E., A comparative study by means of the Rorschach method of personality development in twenty pairs of identical twins, Genet. Psychol. Mono., 1938, 20, 461-556

51. Vernon, P. E., Rorschach inkblot test, II, Brit. J. Med. Psychol., 1933, 13, 179-205

52. _____, The matching method applied to investigations of personality, Psych. Bull, 1936, 33, 149-177

53. Wertham, F., and Bleuler, M., Inconstancy of the formal structure of the personality; experimental study of of the influence of mescaline on the Rorschach test, Arch. Neurol. Psychiat., 1932, 28, 52-70