EMOTIONAL TONE OF PRETEST ACTIVITY AND RETEST RELIABILITY OF THE GOODENOUGH DRAW A MAN TEST

by Ann M. McCormack

Thesis presented to the School of Psychology and Education of the University of Ottawa as partial fulfillment of the requirements for the degree of Master of Arts

Ottawa, Ontario, 1965

ACKNOWLEDGMENTS

This thesis was prepared under the supervision of Mr. J. Gilles Chagnon, M. Ps., Assistant Professor, of the School of Psychology and Education, University of Ottawa.

The author is indebted to Sister Mary Andrew Hartmann, Ph. D., of the Referral Services Division, Ottawa Separate School Board, for her assistance in carrying out this research in the schools, and to the Separate School Board for giving permission to do so.

I also wish to thank Miss Donna Harrison for her assistance in the collection of the data.

CURRICULUM STUDIORUM

Ann Meredith McCormack was born July 6, 1933, in Wilmington, Delaware. She received the Bachelor of Arts degree in Psychology from the University of Delaware in June, 1954, and a Certificate in Education from the University of Manitoba in May, 1961.

TABLE OF CONTENTS

Chapter		page
	INTRODUCTION	vi
I.-	REVIEW OF THE LITERATURE	1
	1. Reliability of the Scale	1
	2. Affective Involvement and Reliability of the Scale	7
	3. Summary and General Hypotheses	17
II.-	EXPERIMENTAL DESIGN	18
	1. Method	18
	A. Structure of the Design	18
	B. Task	19
	C. Instrument and Subjects	20
	D. Assignment of Subjects to Groups	21
	E. Specific Research Hypotheses	23
	2. Procedure	23
	3. Scoring of Drawings	26
	4. Statistical Procedures	27
III.-	PRESENTATION AND DISCUSSION OF THE RESULTS	30
	1. Results of the Experiment	30
	A. Scorer Self-consistency	30
	B. Comparison of Means	31
	C. Comparison of Correlations	35
	2. Discussion of Results	37
	A. Means	37
	B. Correlations	40
	SUMMARY AND CONCLUSIONS	44
	BIBLIOGRAPHY	46
Appendix		
1.	SAMPLE COMPOSITION FORMS	48
2.	RAW SCORE DATA	55
3.	ABSTRACT OF <u>Emotional Tone of Pretest Activity and Retest Reliability of the Goodenough Draw A Man Test</u>	62

LIST OF TABLES

Table — page

I.- Means and Standard Deviations for First and Second Administration of the D.A.M. to Five Groups of Subjects 32

II.- Test-retest Reliability Coefficients (Pearson r) and Fisher Transformation Values (z) for Five Subject Groups on the D.A.M. Test 34

III.- Critical Ratio Values used to Test Significance of Difference among Fisher z's 55

IV.- Critical Ratio Values used to Test Significance of Difference among Fisher z's for Combined Groups 56

V.- Means and Standard Deviations for First and Second Administration of the D.A.M. to Five Groups of Subjects (raw scores) 57

VI.- Test-retest Reliability Coefficients (Pearson r) and Fisher Transformation Values (z) for Five Subject Groups on the D.A.M. Test (raw scores) . 58

VII.- Critical Ratio Values used to Test Significance of Difference among Fisher z's (raw scores) . . 59

INTRODUCTION

The reliability of a measuring instrument is as important to the psychologist as it is to the scientist in any field. In psychology we can seldom achieve the precision of measurement found elsewhere, due to the nature of our instruments and the nature of what we are measuring. Levels of reliability of our instruments which seem quite satisfactory to us would be totally unacceptable in fields where great precision of measurement is so common as to be taken for granted.

Sources of variability in psychological measurement may originate within the subject, the circumstances of the test situation, or the instrument itself. An awareness of the factors which contribute to the reliability of measurement of each instrument makes it possible to control these sources of variability, thus improving the accuracy of the measurements.

There are several methods of estimating the reliability of a psychological test. These include split half, alternate form, and test-retest reliability. For many psychological tests all these methods can be used to obtain a comprehensive evaluation of reliability. For some tests only one among the methods seems to yield a satisfactory estimate; for other tests none is satisfactory.

The instrument used in the present study is the Good-enough Draw A Man Test, a measure of mental maturity of children. Each of the means of estimating reliability mentioned above has

been explored with regard to this test, and it appeared until very recently that there was only one satisfactory method.

At the time of the original statistical evaluation of the scale, Goodenough indicated that the split scale method of estimating reliability was found to be inadequate[1]. And it was not until 1963 that an alternate form of the test was published. Thus the reliability research that has been carried out has centered upon the retest method.

The problem investigated in the present study concerns whether or not pretest activity which is emotionally toned in nature significantly increases the variability of performance on the D.A.M. Test, the increased variability being reflected in lowered retest reliability coefficients.

In the review of the literature (Chapter I of this study) is an evaluation of studies which have extended our knowledge of the retest reliability of the D.A.M. Test, varying the length of the interval, the type of subject, and examining reliability over the entire age range of the test. Here also is an evaluation of studies which have explored the relationship between the affective state of the subject and drawing performance.

---

1 Goodenough, Florence L., "A new approach to the measurement of the intelligence of young children", in Paed. Seminary, Vol. 33, No. 2, issue of June 1926, p. 185-211.

The second chapter presents the experimental design, describing the method used, procedures followed, scoring of drawings and statistical procedures. The specific research hypotheses are outlined here.

Chapter III includes the presentation and discussion of the results of the statistical analysis of the data. Following this is a section summarizing the research and stating the conclusions reached.

Sources used include books and journal articles available in print and on microfilm through the libraries of the universities in Ottawa.

CHAPTER I

REVIEW OF THE LITERATURE

Within this chapter will be found first, a summary of the nature and value of individual studies reporting test-retest coefficients of reliability for the Goodenough Draw A Man Test, then an evaluation of some studies which introduce the factor of affective involvement into the drawing situation. The last section will briefly summarize the review of the literature and will present the general hypotheses.

1. Reliability of the Scale

When Florence Goodenough introduced the Draw A Man Test[1] (D.A.M.) into the field of psychological measurement, she presented a comprehensive body of research data showing the development and standardization of the scale. Included in this data was a report of the reliability of the test, derived from one segment of the standardization group. This initial reliability coefficient of .937 (P.E. $\pm$.006), which was based on a sample of 194 Grade 1 pupils tested with a one day interval between administrations, was sufficiently high to suggest that the test seemed a very promising addition to the field of measurement of mental maturity of children.

---

1 Goodenough, Florence L., The Measurement of Intelligence by Drawings, New York, World book Company, 1926, xiv-177 p.

Even with this encouraging start, though, psychologists were left with legitimate and important questions concerning the reliability of the test. These questions involved whether or not it was equally reliable at all age levels for which it was appropriate; the extent to which the coefficients of reliability would be lowered with longer intervals between administrations; how preceding activities would affect the stability of test scores; and whether various sub-groups of individuals (for example, retardates) would perform more or less reliably than "normal" individuals. Examination of subsequent reliability studies shows that each was stimulated by, and attempted to answer one or more of these important questions.

The research reported in the remainder of this section is presented to indicate the degree of reliability which can be expected for the D.A.M. Test with varying intervals, ages and types of subjects.

Smith's[2] focus of interest was to examine the reliability of the D.A.M. Test over the entire range of age for which it was developed. He administered the scale to a very large number of pupils (2600) in Grades 1 to 8, then gave a second administration to the same children later that day. Samples of one hundred cases were drawn from the total number of cases at each age

---

2 Smith, F.O., "What the Goodenough Intelligence Test measures", in Psychological Bulletin, Vol. 34, No. 9, issue of November 1937, p. 760-761.

level, from six to fifteen inclusive. He found that all correlation coefficients (Pearson r) for age groups through fourteen years varied between .91 and .96. For the age group of fifteen and sixteen years the coefficient was .84; however this group was beyond the intended ceiling of the test in terms of age. These data were consistent with Goodenough's findings (both psychologists having used a short interval between administrations), and thus served to extend the reliability research over the entire age range of the test.

In his study of the drawing performance of delinquent boys, Hinrichs[3] combined investigation of the performance of a particular subgroup with the use of the test over a longer interval than is usually chosen. His sample ranged in age from eight-and-a-half to fourteen-and-a-half years for the reliability phase of the study, and was composed of seventeen boys from a State Home. They had been given the D.A.M. from six months to thirty months previously, these earlier tests having been scored by a different examiner. The rank order correlation of scores from first and second administrations of the scale was .65. For a second delinquent group the correlation was .79. The author noted that these correlations were relatively low compared with those reported elsewhere, and suggested that this

---

3 Hinrichs, W.E., "The Goodenough drawing in relation to delinquency and problem behavior", in Archives of Psychology, Vol. 26, No. 175, issue of January 1935, p. 83-92.

was due to the small sample size, as well as to the type of subject used. (That is, delinquents would be expected to show greater variability of performance.)

However there were also factors present in this study which may have served to confound the results, such as the use of the scale beyond its intended ceiling (and although an "extension" was employed, it would be important to know its validity); the extreme breadth of the interval (there was no valid ground for assuming the test is equally reliable at thirty month intervals as at six month intervals); and the use of different examiners for administration and scoring of the tests (McCarthy[4] noted the extensive pre-training necessary for scorers to achieve acceptable levels of inter-scorer consistency). In assessing the contribution of this study to the literature as a whole, it is evident that the statistical results were no more valid than the data from which they were derived, and although they may seem "sensible", they must be verified by research which is far more rigorously controlled.

In examining the drawing performance of another type of sub-group, Yepsen[5] administered the D.A.M. to a group of

---

4 McCarthy, Dorothea, "A study of the reliability of the Goodenough Drawing Test of Intelligence", in Journal of Psychology, Vol. 18, Second Half, issue of October 1944, p. 201-216.

5 Yepsen, L.N., "The reliability of the Goodenough Drawing Test with feeble-minded subjects", in J. educ. Psychol., Vol. 20, No. 6, issue of September 1929, p. 448-451.

thirty-seven feebleminded boys whose ages ranged from nine to eighteen. Since the median M.A., established on the Stanford-Binet, was eight years, it is unlikely that the ceiling of the test was exceeded in this instance. There were three administrations of the D.A.M., with a four day interval between first and second, and second and third administrations. Coefficients of reliability were as follows: Test 1,2 .89; Test 2,3 .91; Test 1,3 .91. Variability rarely exceeded one year, and of the subjects who changed rank, half of their scores increased while half decreased. Thus Yepsen concluded with justification that the D.A.M. was highly reliable for re-administration to feebleminded subjects.

Further investigation into the reliability of the D.A.M. with feebleminded subjects was carried out by Brill[6]. His sample was composed of sixty-five boys whose median test age was eight years and four months. The D.A.M. was administered to the group, readministered after an interval of eighteen days, then given for a third time twenty-five days after the second test. The reliability coefficients found were: Test 1,2 .77; Test 2,3 .80; Test 1,3 .68. Brill remarked that although these coefficients were high, they were lower than those found by Goodenough or by Yepsen. Possibly this was due to the longer

---

6 Brill, M., "The reliability of the Goodenough Draw-A-Man Test and the validity and reliability of an abbreviated scoring method", in J. educ. Psychol., Vol. 26, No. 9, issue of December 1935, p. 701-708.

interval used by Brill and, relative to Yepsen's study, the larger number of subjects. He concluded that the D.A.M. could be used with a high degree of reliability with feebleminded boys, from three to six weeks after the original administration.

The lowest correlation obtained in a reliability study using the Goodenough Test was that found by McHugh[7]. He administered the D.A.M. to eighty-three Kindergarten children two weeks before entrance to Kindergarten, and again one to three months later. The mean score as measured by M.A. rose from 58.8 to 65.4 months from first to second administration, with a test-retest correlation coefficient of .51 between the two sets of scores. It is likely that the critical factor in this study was initial school experience; thus the low correlation coefficient obtained would result from all the related factors which contribute toward increasing the difference between drawings done before starting school and those done after a period of school experience.

A study examining the contribution of time to variability of performance on the D.A.M. was done by McCurdy[8]. He tested fifty-nine Grade 1 children twice, with an eleven week interval

---

7 McHugh, G., "Changes in Goodenough I. Q. at the public school kindergarten level", in J. educ Psychol., Vol. 36, No. 1, issue of January 1945, p. 17-30.

8 McCurdy, H.G., "Group and individual variability on the Goodenough Draw-A-Person Test", in J. educ. Psychol., Vol. 38, No. 7, issue of November 1947, p. 428-436.

between administrations. The correlation coefficient between the two sets of scores was .69. Evaluating the reliability studies done prior to his own, McCurdy concluded that variability of performance on the D.A.M. is to an important degree a function of time.

There are other factors as well which increase variability of performance on the D.A.M. Test, one of which will be discussed in the following section.

2. Affective Involvement and Reliability of the Scale

The research presented in this section will center on studies which have explored the relationship between the affective state of the subject and drawing performance.

Thus far there has been no mention of the "projective hypothesis"[9] but its inclusion in the literature seems appropriate at this point. According to Frank, the individual's personality operates on the environment in such a way that he imposes his "private world" of meanings upon the stimulus situation. In the true sense of the projective use of drawings the emphasis is upon a qualitative evaluation of their contents, inferring from this the personality characteristics (including the affective component) of the subject.

---

9 Frank, L.K., "Projective methods for the study of personality", in Journal of Psychology, Vol. 8, Second Half, issue of October 1939, p. 389-413.

However there are studies which suggest that the individual's affective state can be revealed by quantitative analysis of drawings as well. This use of drawings is not within the true meaning of the projective hypothesis since Frank's intention was to break away from the quantitative tradition in psychology which he felt was not adequate to study the "private world" of personality.

Thus the bulk of studies using human figure drawings as a projective device need not be considered here; rather the research centered upon the individual's affective state as it can be measured objectively in drawing performance will be the main concern. These investigations tend to focus upon the effect of some activity or state upon individual or group mean scores, and unfortunately do not concern themselves with retest reliability for the most part. Nevertheless it will be helpful to consider such studies here in order to clarify the role of affective involvement in drawing performance.

The following two studies, which are concerned with the effect of physical activity upon subsequent drawings, are of secondary interest and will be discussed very briefly. In Mott's[10] research the children were asked to draw a man, then to exercise various body parts verbalizing their activity as

---

10 Mott, Sina M., "Muscular activity an aid in concept formation", in Child Development, Vol. 16, 1945, p. 97-109, and reported in D.B. Harris, Children's Drawings as Measures of Intellectual Maturity, New York, Harcourt, Brace & World, 1963, p. 23.

they did so, then to make a second drawing. The drawings following the exercise showed increased scores for the exercised parts, which were found to be represented with more care.

Harris[11] gave a fifteen minute exercise period each day for two weeks, followed by the D.A.M. for one group of Grade 1 children, while another group did the D.A.M. every day for two weeks without exercising first. He found no significant increase in arm or leg emphasis, or in motion of the drawings of the exercise group, and suggested that the verbal involvement of Mott's subjects may have accounted for the differing results of the two studies.

Turning now to research which considers the influence of a gratifying experience upon D.A.M. performance, the following study examines the effect of gratification upon a more general type of drawing activity.

Reichenberg[12] presented children with a repetitive line-drawing task which they performed until they refused to continue any longer. At the point of refusal one group was given a "joyful experience", while the other was not. Then all were given the task once again. The author reported improved quality

---

11 Harris, D.B., "Intra-individual vs. inter-individual consistency in children's drawings of a man", in American Psychologist, Vol. 5, No. 7, issue of July 1950, p. 295; reported more fully in Harris' book, op. cit., p. 93-95.

12 Reichenberg, W., "An experimental investigation of the effect of gratification upon effort and orientation to reality", in American Journal of Orthopsychiatry, Vol. 9, No. 1, issue of January 1939, p. 186-203.

and quantity of work following the joyful experience. Unfortunately, the results were not presented in a manner which indicated the statistical significance of the differences found. But in spite of this drawback, the study merits a place in the literature for its demonstration that drawing performance following an affective experience can show quantitative and qualitative changes in the predicted direction.

In a later study Reichenberg-Hackett[13] used the D.A.M. Test to examine the effect of a gratifying experience on drawing performance. Since statistical tests of significance were applied in this instance, and positive results were obtained, the study will be examined in closer detail. The experimental group consisted of 106 children whose mean age was ten years and six months. The control group was composed of one hundred children whose mean age was nine years and ten months. The controls were given the D.A.M., then the Stanford-Binet as an interim activity, then the D.A.M. again. The experimental group was divided into three sub-groups, one having the D.A.M. followed by a puzzle box, then the D.A.M. again. The puzzle box involved the child's manipulation of a box to get a small gift, accompanied by warm interaction on the part of the examiner. A second sub-group did the D.A.M., then received a gift (without experimenter-interaction), then was given the D.A.M. again. The third sub-group

---

13 Reichenberg-Hackett, W., "Changes in Goodenough drawings after a gratifying experience", in American Journal of Orthopsychiatry, Vol. 23, No. 3, issue of July 1953, p. 501-517.

was given the Cloud Picture Test for ten minutes as a neutral activity between D.A.M. administrations.

In considering the results of this study, two factors are immediately obvious: 1) the control group was poorly chosen, since the D.A.M. IQ scores dropped from 103 on first administration to ninety-six on second administration, an occurrence which was very likely due to the taxing nature of the intervening activity; 2) the length of the interval was not the same for all groups.

In comparing the retest mean scores of the experimental groups with that of the control group, the puzzle box group showed an increase of five IQ points, whereas the scores of the other two experimental groups decreased two points. The control group scores fell seven points; thus comparing mean scores across groups Reichenberg-Hackett found these differences to be statistically significant (the level of significance was not indicated). However with a more appropriate control group, it is likely that the scores of the latter two experimental groups would not significantly differ from those of the control group.

Within the puzzle box group there was a decided change in mean score in the predicted direction. The IQ scores of this group increased from 104 to 109, while scores of the "gift" group dropped from 102 to 100, and the "cloud picture" group scores fell from 104 to 102. Observation of the children in the gratifying encounter showed that their energy output

increased and they were able to draw upon more resources as a result of the encounter. The author concluded that

> . . . positively toned affective states influenced the work of the children in our 'puzzle box group' in the direction of higher achievement as indicated by higher scores on the Goodenough scale . . .
>
> That this acceleration of ability was due to affective factors stands out when we compare the puzzle box data with the data of the other groups. In the puzzle box situation a combination of gratifying factors was brought to bear on the situation. The child enjoyed success, attributable to the self, plus a concrete reward plus excellent social relations. The receiving of a concrete gift alone did not influence achievement on the Goodenough score to the same extent for the second group. This suggests that to children in our age range the experience of ego-involving success is more meaningful in terms of gratification and change of mood than a mere 'reward'. . . . Results also indicate that the marked variability of performance over short intervals is due to the effect of the immediately preceding activity of affective states and of mental content. This point, suggested by McCarthy, is substantiated by our findings.[14]

In the area of deficiency, Hunt and Patterson[15] investigated the effect of gratification upon drawing performance of familially mentally deficient children, using as subjects fifty boys whose ages ranged from seven to fourteen, and whose I.Q.s ranged from thirty to eighty. Two sets of four administrations each of the D.A.M. were given, with a thirty-one day interval between the first and second sets. The final trial of

---

14 Ibid, p. 513-514.

15 Hunt, E. and R.C. Patterson, "Performance of familial mentally deficient children in response to motivation on the Goodenough Draw-A-Man Test", in American Journal of Mental Deficiency, Vol. 62, No. 2, issue of September 1957, p. 526-529.

the first set was accompanied by a candy; in the second set a candy plus verbal urging was provided. Only one significant difference between scores within the sets was found ($p = .05$), that being under the condition of verbal reinforcement. Reliability coefficients were reported only incidentally: ". . . all correlation coefficients were above .60."[16] This would seem to be low in view of the short interval between tests within each set, and the nature of the sample, which should also add stability to the performance.

A well designed, carefully executed study of the reliability of the D.A.M. Test was reported by McCarthy.[17] Her threefold intention was to determine

> . . . how consistent the same scorers are with themselves in scoring identical drawings independently on two occasions; . . . how well scorers agree with each other on the scoring of identical drawings; . . . the relationship between retests of the same children over a short interval when both sets of tests are scored by the same person.[18]

The Grade 3 and 4 pupils from one school ($N = 386$) were given two administrations of the D.A.M. with an interval of one week between administrations. Scorers were given a lengthy period of training until a high level of agreement regarding the criteria of scoring was reached (average inter-scorer agreement was .90).

---

16 Ibid, p. 327.

17 McCarthy, op. cit.

18 Ibid, p. 203-204.

The results of McCarthy's research are of interest in several areas:

1) Mean IQ score for the first administration was 100.3; for the second administration it was 95.8. There was no statement regarding the statistical significance of this difference; however, for a sample of this size it is a substantial difference and may well approach significance. The drawings were scored three times; a comparable difference was found within the second, but not within the third scoring.

2) Retest reliability coefficients were reported by scorer, each of whom scored the drawings from both administrations. The mean Pearson r was .68. McCarthy commented that this level of reliability was lower than that found by Goodenough, Yepsen or Smith, and seemed to resemble more closely that of Hinrichs, who used much smaller samples, varying scorers, and greatly increased intervals. Indeed, this seems an astonishing parallel to draw since the studies differ so greatly in adequacy of design and care invested in controlling significant variables.

3) Scorer self-agreement[19] was reported as .95, .92 and .93 for the three scorers. These coefficients were consistent with those reported by other investigators.

4) Regarding group variability, McCarthy noted that on the second administration, two-thirds of the sample fell between

---

19 Evidently based on comparisons of total scores.

± 5.06 points of their scores from the first administration, thus producing a year or more change in M.A., which she felt was quite large.

Thus there were several areas of concern in her results: the change in mean IQ scores; the unexpectedly low retest reliability coefficients, and consequent high variability in scores from first to second administration. Noting that the group activity before the first administration had been a positively toned written task, and before the second administration it had been a negatively toned written task, McCarthy stated that

> . . . it is possible that some children were somewhat emotionally elated at the time of the first drawing and emotionally disturbed from the recent recollection of an unpleasant event at the time of the second drawing. This difference in emotional setting might have made them less attentive and more careless in the execution of the second drawing. It would be interesting to study further the effect of preceding activities and mental content, particularly the effect of various affective states on drawing and other performances.[20]

It was this possibility, that the affective state of the subjects prior to drawing activity may have increased the variability of performance, thus lowering the retest reliability coefficients, which led to the formulation of the design of the present study. An indication of the plausibility of this suspicion is found in Lindquist's discussion of the sources of error of measurement (hence variability of performance), as he

---

20 McCarthy, op. cit., p. 206.

includes "the fluctuating character of the individual's mental, emotional, or physiological state."[21]

It seemed important, in devising a suitable design to study the mean scores and reliability coefficients resulting from differently emotionally toned pretest activity, to consider not only presence or absence (relative) of affective state in the various groups of subjects, but the type of affective state involved as well. The questions to be answered, stemming from McCarthy's study, were: 1) did an emotionally toned task preceding the drawings contribute substantially to the low coefficients of reliability found and, if so, 2) was the change in the nature of the emotional tone of the tasks on the two administrations the important factor, or simply the presence of affective involvement?

To answer the first question, groups would be needed having emotionally toned tasks which were substantially the same before the two drawing performances, as well as a group whose pretest activity was not emotionally toned. Answering the second question would require additional groups which were presented with emotionally toned tasks whose nature changed from first to second administration.

---

21 Lindquist, E.F., A First Course in Statistics, New York, Houghton Mifflin Company, 1942, p. 215.

5. Summary and General Hypotheses

In summary, the review of the literature has included first, the findings of the research which focused on determining the retest reliability of the D.A.M. Test, varying factors such as length of interval, age, and type of subject.

Following this was an evaluation of research which considered the affective state of the subject at the time of, or just prior to the drawing performance. McCarthy's study was considered in some detail since the present research followed directly from it.

The purpose of the present study is to verify and extend existing knowledge about the influence of affective state in children on their performance on the Goodenough Draw a Man Test. Previous findings suggest that mean scores of the group having first a pleasant then an unpleasant task will fall; that the opposite will occur for the group experiencing the reverse conditions; and that no difference will be observed in the mean scores of the other groups of subjects. It also seems likely that the groups having the affectively toned tasks will show lower test-retest coefficients of reliability than the group whose pretest activity is not affectively toned.

In the following chapter will be found the specific research hypotheses and a description of the experimental design which was devised to test these hypotheses.

CHAPTER II

EXPERIMENTAL DESIGN

The purpose of this chapter is to describe the method used to test the general hypotheses presented in the previous chapter. There are four sections in the chapter as follows: 1) method; 2) procedure; 3) scoring of drawings; 4) statistical procedures used in testing the experimental hypotheses. The specific research hypotheses are stated at the end of the first section.

1. Method

A. Structure of the Design

This design involves five subject groups of approximately fifty children each, three pretest tasks, the Draw A Man Test, and an interval of one week between test and retest. The sequence of activity followed by the five groups was structured as follows:

Group I Task A--D.A.M.----1 week----Task A--D.A.M.

Group II Task B--D.A.M.----1 week----Task B--D.A.M.

Group III Task C--D.A.M.----1 week----Task C--D.A.M.

Group IV Task A--D.A.M.----1 week----Task B--D.A.M.

Group V Task B--D.A.M.----1 week----Task A--D.A.M.

An investigation of all the combinations of the three tasks with each other would involve nine groups; however, it

is felt that the above five groups will fulfill the purpose of this investigation and provide the data necessary to test the hypotheses.

B. Task

The two pretest activities which occurred accidentally during McCarthy's study are used purposefully in the present research, with a third task added. Equivalent forms of the tasks were developed to avoid the less involving repetition inherent in Groups I, II and III on second administration. These groups were divided so that half of each group was given one form of the task on first administration and the alternate form on second administration. The order was reversed for the other half of each group. For example, half of Group I was given Task A_a followed by A_b, while the other half began with A_b and finished with A_a.

Task A. - Task A instructed the subjects as follows:

> You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.
> THE BEST THING THAT EVER HAPPENED TO ME

The topic for the alternate form of Task A was THE HAPPIEST DAY I CAN REMEMBER. These topics were felt to elicit a pleasant emotional tone.

Task B. - In Task B the subjects were instructed:

> You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.
> THE WORST THING THAT EVER HAPPENED TO ME

The alternate form of Task B presented the topic THE SADDEST DAY I CAN REMEMBER. These topics were felt to elicit an unpleasant emotional tone.

Task C. - Task C gave the instructions:

> You will have 15 minutes to write all the words you can think of beginning with the letters shown. Go ahead and begin. 1. How many words can you think of beginning with the letters B or R? 2. How many words can you think of beginning with the letters D or S?

The alternate form of Task C used identical instructions, substituting the letters C, P, M and T. This was not intended as a power task; hence the letters selected were letters offering the largest number of words to choose from, according to the dictionary consulted.[1] The purpose of this activity was to engage the subjects in a written task which would elicit no consistent emotional tone across the group. For this reason it is referred to as the "neutral" task or activity. Sample sheets of each task are presented in Appendix 1.

C. Instrument and Subjects

The instrument used as a measure of drawing performance was the Goodenough Draw A Man Test. The instructions were

---

1 The Pocket Oxford Dictionary of Current English, compiled by F.G. Fowler and H.W. Fowler, London, Oxford University Press, 1957, p. xvi-960.

followed as outlined in Goodenough's book[2], and were combined with task instructions as noted in the section immediately following on Procedure. Goodenough's scoring system was also adhered to.

The subjects were 313 children in ten Grade 4 classrooms in the English speaking Separate Schools of Ottawa. Nine schools were used (one having two full sections of Grade 4), representing several socioeconomic levels and varying geographical areas of the city. Thirty-one children were dropped after completion of testing because of absence from either the first or second administration. Twenty-one drawings were discarded because they were unscorable. These fell into several categories: some figures had been drawn and erased; others were drawn disproportionately large so that they ran off the page; in the majority of cases a head, or head and trunk were drawn with considerable detail but the figures had not progressed beyond this in the time allowed. Thus there remained 261 usable pairs of drawings, 139 done by boys and 122 by girls.

D. Assignment of Subjects to Groups

By having all groups represented in each classroom, it was hoped to avoid the type of bias which may occur when classrooms are used as units.

---

2 Goodenough, Florence L., Measurement of Intelligence by Drawings, New York, World Book Co., 1926, p. 85.

The method of assigning subjects to groups which was originally planned was to use a table of random numbers. This was tried for a small classroom, but it was evident that the groups could, by chance, be very unevenly represented within one classroom. In this instance, nearly fifty percent of the class fell into Group V, while none were in Groups IIa or IV. Since there was a large variation in the number of pupils across classrooms, and because of the difference in socioeconomic level (hence possibly mean I_q level) among the schools, it seemed preferable to force an even representation of groups within each classroom.

To effect this even representation, five sets of ten tags each were made up, each set being composed of tags marked Ia, Ib, IIa, IIb, IIIa, IIIb, IV, IV, V, V, indicating the groups. For a classroom of twenty to twenty-nine pupils, three sets of tags were placed in a container, shaken, and tags were drawn out singly until the necessary number had been obtained. For a class of thirty to thirty-nine, four sets were used, and for a class of forty to forty-nine, all five sets were used.

The composition forms were arranged in the order determined by the tags, and they were distributed in the classrooms in this order, down each row of desks. In this manner, each child was automatically assigned to a group as he received a composition form.

Since the group assignment of each child was thus fixed on first administration, the second set of papers was distributed

with the children's names on the papers. Each pair of compositions and drawings was later coded to indicate the group number, school, examiner, test or retest, and sex of the child.

E. Specific Research Hypotheses

It has been suggested that drawing performance may be influenced by affectively toned pretest activity. In order to obtain experimental verification of the general hypotheses presented in Chapter I, the following specific hypotheses were formulated and are stated in null form:

1. There are no significant differences among test and retest means on the D.A.M. Test within any of five groups of subjects given different combinations of neutral, positively and negatively toned pretest activity;

2. There are no significant differences among test and retest means on the D.A.M. Test across five groups of subjects given different combinations of neutral, positively and negatively toned pretest activity;

3. There are no significant differences among retest reliability coefficients on the D.A.M. Test among five groups of subjects given different combinations of neutral, positively and negatively toned pretest activity.

The specific procedures used to test these hypotheses are outlined in the following sections.

2. Procedure

Contact with the schools was initiated through the Referral Services Division of the Ottawa Separate School Board,

and arrangements for testing in the classrooms were made.

Two examiners participated in collection of the data, each testing five classrooms on two occasions. With respect to the use of two examiners, Harris reports that in a study designed to investigate this variable, the

> . . . person of the administrator had very little influence on either the mean score achieved by the class or on the rank order of children's scores within the class. In none of the classes was the difference between the two administrators statistically significant.[3]

Testing was carried out in all classrooms in the mornings, immediately following opening exercises. It took approximately thirty minutes on each occasion. The examiners were presented to the group by name by the teacher, and read the following instructions, identical for all groups:

> Good morning boys and girls. I am very happy to be visiting your classroom this morning.
>
> I am going to pass out some papers to you. Each paper will have directions printed on it, telling you what to do. Take the top paper and pass the rest along. As soon as you get your paper, fill in your name, birthdate and age in the blank spaces at the top. Then go ahead and begin. You will have 15 minutes to do what it asks. Boys and girls in many other schools will be doing the same things you are doing, so each one do the very best you can. (after 10 minutes)
>
> You have 5 minutes to finish up. (after 5 more minutes) Everyone should be finished now. Make sure your name and birthdate are on the paper. Pass your paper to this end of the row. (collect papers and distribute blank paper and pencils)
>
> Each of you take one blank sheet of paper and a pencil. Write your name and birthdate in the upper

---

3 Harris, D., Children's Drawings as Measures of Intellectual Maturity, New York, Harcourt, Brace & World, 1963, p. 92.

left hand corner (hold up paper and point). Now turn the paper over to the other side.

On these papers I want you to make a picture of a man. Make the very best picture that you can. Take your time and work very carefully. I want to see whether the boys and girls in ----- school can do as well as those in other schools. Try very hard and see what good pictures you can make. Make the whole man. (after 10 minutes)

Finish up now. Make sure your name is on the back of the paper. Pass your paper and pencil to the end of the row. Thank you boys and girls; that's all.

Materials supplied during testing included mimeographed forms for the written tasks (see Appendix 1 for samples), blank sheets of white 8½ x 11 bond paper for the figure drawings, and pencils of uniform medium quality lead, freshly sharpened before use by succeeding groups of pupils.

Following the first administration the teacher was told that another visit would be made in one week's time; however, no indication was given as to the nature of the second visit, or the purpose of the study. Each teacher was asked not to discuss the compositions or drawings with the children until the end of the second visit.

There was a good deal of confusion among the children about their birthdates, and teachers' registers were used to check doubtful cases.

On the occasion of the second administration, the instructions were somewhat modified as follows:

Good morning boys and girls. I am glad to be visiting you again today.

I'm going to hand out some papers and I want you to read the directions as you did before and do what

they ask. I have looked over the things that you wrote last week, and they are very good. Each of you do the best that you can again today.

Everyone who was here last week will get a paper with his name at the top. Write your name again in the blank space, along with your birthdate and age. Then go right ahead and begin. You should not have to ask questions this time. Anyone who was not here last week please raise your hand (pass out papers) (after 10 minutes)

You have 5 minutes to finish up. (after 5 more minutes) Everyone should be finished now. Make sure your name and birthdate are on the paper. Pass your paper to the front of the row. (collect papers and distribute blank paper and pencils)

Each of you take one blank sheet of paper and a pencil. Write your name and birthdate in the upper left hand corner (demonstrate). Now turn the papers over to the other side. Once again, on these papers I want you to make a picture of a man. Make the very best picture that you can. Take your time and work very carefully. I want to see whether the boys and girls in ----- school can do as well as those in other schools. Try very hard and see what good pictures you can make. Make the whole man. (after 10 minutes)

Finish up now. Make sure your name is on the back of the paper. Pass your paper and pencil to the end of the row. Thank you boys and girls.

Questions were answered in a nondirective manner, except when they concerned using crayons to color the drawings, which was not permitted. In the case of multiple drawings, the best one was chosen for scoring, as suggested by Goodenough.[4]

5. Scoring of Drawings

The scoring system outlined by Goodenough was closely followed. After the coding of the drawings was completed, they

---

4 Goodenough, op. cit., p. 89.

were thoroughly shuffled and all were scored a first time. If drawings were encountered which had remained in pairs, or within recognizable proximity, the second was removed a good distance from the first so that scoring would be unbiased. Data sheets were divided twice, first into five groups, then according to test or retest scores, to facilitate handling of the data.

After completion of the initial scoring, every tenth drawing was re-scored in order to obtain a total point estimate of scorer self-consistency (N = 108). After this was done these drawings were numbered 1 to 108. A tag representing each was placed in a container and fifty were drawn out. This latter group of drawings was used to estimate scorer self-consistency by means of a point-by-point analysis of the fifty-one Goodenough scoring criteria.

4. Statistical Procedures

This study was designed to examine the effect of pretest activity on test-retest reliability coefficients, and on group mean scores. Computations were initiated using raw-score data (see Appendix 2), then raw scores were converted into IQ scores. These were recorded on Dayhaw Correlation Sheets, whose formulas were followed in obtaining Pearson product-moment correlations, means, and standard deviations.

Scorer self-consistency was established by obtaining a Pearson product-moment correlation value based on raw scores on

first and second scorings of 100 drawings. The self-consistency rating for fifty drawings compared on a point-by-point evaluation was computed by means of percentage of agreement between first and second scoring of each drawing.

To test the means within groups to establish the level of significance of the differences between these pairs of scores, Guilford's formula for determining the standard error of a difference in correlated data was used.[5] The difference was then divided by the standard error of the difference, yielding a critical ratio value in the usual manner.

To test the means across the groups for significance of difference, Guilford's formula for obtaining the standard error of a difference between uncorrelated means was used[6], followed by determination of a critical ratio value.

To test significance of differences among retest reliability coefficients of the five groups, each coefficient was converted into a Fisher *z* coefficient.[7] Standard errors of the *z* differences were established[8], and critical ratios computed.

---

5 Guilford, J.P., Fundamental Statistics in Psychology and Education, New York, McGraw-Hill, 1956, p. 136.

6 Ibid, p. 183.

7 Ibid, p. 194.

8 Guilford, loc. cit.

The four sections of this chapter have covered the experimental design of the study. The next chapter contains the presentation and discussion of the results.

CHAPTER III

PRESENTATION AND DISCUSSION OF THE RESULTS

This chapter contains the results of the statistical treatment of the data, and a discussion of those results.

1. Results of the Experiment

The raw scores derived from the Goodenough point scale evaluations of the drawings were converted into M.A., then into I.Q. scores, and the statistical analysis in all cases except scorer self-consistency was done using I.Q. scores. In the latter instance raw scores were used.

A. Scorer Self-consistency

There were two estimates of scorer self-consistency. First, 108 drawings were re-scored, and the two sets of total raw score points were correlated, yielding a Pearson r of .923.

Then the scorings of fifty of these 108 drawings were compared in a point-by-point analysis. The mean scorer percentage of agreement for this more rigorous estimate was 89.6%. These levels of scorer self-consistency are consistent with those reported in other studies, and are considered satisfactory.

B. Comparison of Means

Table I presents test and retest mean scores for the five groups of subjects in the experiment. To test the significance of the differences obtained between the pairs of means, Guilford's formula for computing the standard error of a difference between correlated means was used, along with the critical ratio.

None of the pairs of mean IQ values of the groups was found to have changed significantly from test to retest. Thus the expected relationship between type of pretest activity and mean drawing performance was not reflected in the data, and the null hypothesis cannot be rejected.

Nor was any anticipated trend of mean values observed across groups. It was expected that mean scores would be higher for groups having the positively toned task, and lower for those having the negatively toned task. Inspection of the data reveals the opposite trend, and it was not felt necessary to do tests of significance except between Groups I and II. The standard error of difference between uncorrelated means was calculated, and the critical ratio of 1.40 was not significant. Here again, the null hypothesis cannot be rejected.

To establish the equivalence of the two forms of each task used in Groups I, II and III, 't' tests were employed, and no significant differences were found. Thus the two forms of each task were considered equivalent.

Table I.-

Means and Standard Deviations for First and Second Administrations of the D.A.M. to Five Groups of Subjects[a]

Group	N	M_1	s_1	M_2	s_2
I	51	95.42	19.14	95.65	16.44
II	56	100.60	17.96	98.60	15.88
III	51	96.30	17.20	96.90	15.79
IV	54	96.45	15.16	97.30	16.95
V	49	97.60	14.47	98.20	14.50

a Calculations based on I_2 scores

C. Comparison of Correlations

In order to test the significance of differences among the five Pearson *r*'s, each correlation coefficient was transformed into a Fisher *z* value, as shown in Table XI.

The standard error of the difference between the pairs of *z* values was then computed, and critical ratios were established to test the significance of the difference between each pair of values. The critical ratios are presented in Table XII.

Examination of the critical ratio values revealed that in the comparison of Groups III and IV, the difference between the Fisher *z*'s was significant at $p < .05$. Thus the null hypothesis can be rejected.

It was observed that the correlations were aligned on a continuum according to type of treatment; that is, the highest correlation was obtained for Group III (neutral activity), the next highest for Groups I and II (emotionally toned activity), and the lowest for Groups IV and V (change of emotional tone of activity). Since the differences between I and II, and IV and V were not significant, the data in these pairs of groups were combined and tested for significance, as shown in Table IV.

On the basis of the combined groups, a significant difference was found between Groups III and IV-V ($p < .05$).

Table II.-

Test-retest Reliability Coefficients (Pearson r) and Fisher Transformation Values (z) for Five Subject Groups on the D.A.S. Test

Group	N	Pearson r	Pearson r (combined gps)	Fisher z	Fisher z (combined gps)
I	51	.82	.78	1.16	1.05
II	56	.80		1.10	
III	51	.86		1.29	
IV	54	.70	.73	.87	.93
V	49	.77		1.02	

Table III.-

Critical Ratio Values used to Test Significance of Differences among Fisher z's

Group	I z=1.16	II z=1.10	III z=1.29	IV z=.87
II z=1.10	.30			
III z=1.29	.63	.96		
IV z=.87	1.45	1.18	2.10[a]	
V z=1.02	.69	.40	1.32	.74

a Significant at $p < .05$

Table IV.-

Critical Ratio Values used to Test Significance of Differences among Fisher *z*'s for Combined Groups

Group	I-II *z*=1.05	III *z*=1.29
III *z*=1.29	1.41	
IV-V *z*=.93	.85	2.11[a]

a Significant at $p < .03$

2. Discussion of Results

The statistical findings will be discussed here relative to the findings of other studies.

A. Means

Failure of the means to show a significant directional change from test to retest for Groups IV and V, or across groups for test or retest suggests that if any change in performance occurred as a result of varying the pretest tasks, it was not reflected in a consistent or predictable rising or falling of the mean scores of the five groups of subjects.

Other studies have reported conflicting findings regarding test and retest means. Among those studies not involving affective states, McCurdy[1] observed a tendency for retest mean values to increase (interval of eleven weeks), whereas Brill[2] found a trend toward decreased scores on successive administrations (interval of six weeks) which was not statistically significant however. McHugh[3] found gains over a three month period

---

1 McCurdy, H.G., "Group and individual variability on the Goodenough Draw A Person Test", J. educ. Psychol., Vol. 38, No. 7, issue of November 1947, p. 425-430.

2 Brill, M., "The reliability of the Goodenough Draw-A-Man Test and the validity and reliability of an abbreviated scoring method", in J. educ. Psychol., Vol. 26, No. 9, issue of December 1935, p. 701-708.

3 McHugh, G., "Changes in Goodenough I.Q. at the public school kindergarten level", in J. educ. Psychol., Vol. 36, No. 1, issue of January 1945, p. 17-30.

which were highly significant; however in an earlier discussion of McHugh's study it was noted that school experience appeared to be the critical factor involved here. Such pronounced gains are atypical of the total range of the D.A.M. Test.

Among studies involving affective states, Reichenberg-Hackett[4] did not report tests of significance to compare mean performance within his groups. Inspection of the data presented makes it seem unlikely that there would be a significant difference. However there was a mean increase of five I.Q. points on retest for the group given greatest gratification, while a slight decrease was evident for the other groups.

McCarthy[5] reported a decrease of about five I.Q. points for retest mean scores and suggested that the lower scores on retest were due either to less motivation on repetition of the task (which seems unlikely on the basis of the above mentioned studies), or to the influence of the pretest affective involvement. In her study all drawings were scored three times, and since this five point difference between the mean values appears in only two of the three scorings, its stability might be questioned.

---

4 Reichenberg-Hackett, W., "Changes in Goodenough drawings after a gratifying experience", in Amer. J. Orthopsychiat., Vol. 23, No. 3, issue of July 1953, p. 501-517.

5 McCarthy, Dorothea, "A study of the reliability of the Goodenough Drawing Test of Intelligence", in J. Psychol., Vol. 18, Second Half, issue of October 1944, p. 201-210.

To summarize the above findings, it appears that when no extraordinary factors are operating, the mean group scores tend to remain the same on retest; that is, they do not show a consistent increase (practice effect) or decrease (lessening of motivation).

When affective factors appear, the scores tend to rise or fall in the expected direction; that is, they may increase when gratification is present and decrease under conditions of pleasant followed by unpleasant emotionally toned tasks.

Mean scores in the present study however failed to show other than chance fluctuations. This was unexpected since the pretest activity was highly similar to that used in the McCarthy study. However, if the decrease in mean scores in her study is not consistent over three independent scorings of the drawings, it may have been a chance occurrence that her scores showed a change in the expected direction in two of the scorings. If this is the case, then it is likely that a pencil and paper task given in a group situation is not sufficient to evoke an intense and consistent emotional tone over a group of subjects to the extent of being reflected in predicted directional changes of mean scores. Such an effect would more likely appear following individual testing of subjects such as that done by Reichenberg-Hackett, or following a more intensely involving group situation such as presentation of a film with a very strong emotional tone.

B. Correlations

Of the five Pearson r's computed using test and retest scores of the five groups of subjects, the highest correlation is noted for Group III, whose pretest activity has been described as neutral. As affective involvement is employed in the other four groups, an interesting trend is noted as the correlations fall along a continuum, with the highest being shown for the neutral group (Group III), the next highest for the groups having the same type of affective involvement on test and retest (Groups I and II), and the lowest being evident in the groups having a change in affective state from test to retest (Groups IV and V). The only significant difference among these correlations was between Groups III and IV ($p < .05$); however, when the data for Groups IV and V were combined, the difference between the combined correlation (IV-V) and the correlation for Group III was also significant, this time at $p < .05$.

Also of interest when considering variability of scores is the variability of the groups as entities on both occasions, as reflected in the standard deviations. Inspection of the standard deviations yields one factor of interest: in the first three groups the variability of each group is less on second administration than on first administration. That is, individual performances tend to deviate less on the whole from the average performance on retest. That this is the usual finding

can be seen in several studies (among them those of McCurdy[6], Iepsen[7], and McHugh[8]) reporting less variability in performance on second administration of the D.A.M. Test. Brill states that

> The distribution of the second administration tends to greater compactness than that of the first one . . .[9]

Shanan interprets the drop in variability from session to session as giving support to the notion that "repetition of task will lower variability"[10]; thus "It appears as if a person becomes more consistent on a task after repeating it."[11]

In contrast to the usual decrease of standard deviations seen in Groups I, II and III, Groups IV and V exhibit an increased standard deviation on retest, reflecting a greater average spread between the scores on second administration for these groups. Although the increases are not large, they are noteworthy because they occur in the groups whose combined consistency of

---

6 McCurdy, op. cit.

7 Iepsen, L.N., "The reliability of the Goodenough Drawing Test with feebleminded subjects", in J. educ. Psychol., Vol. 20, No. 6, issue of September 1929, p. 448-451.

8 McHugh, op. cit.

9 Brill, op. cit., p. 704.

10 Shanan, J., "Intraindividual response variability in figure drawing tasks", in Journal of Projective Techniques, Vol. 26, No. 1, issue of March 1962, p. 107.

11 Ibid, p. 108.

performance over two occasions was significantly lower than that of the neutral group.

Returning now to McCarthy's study, the points of similarity between her research and this study will first be noted. Similarities of design include: 1) the same pretest activity for one group of this study as for her entire sample; 2) the same measure of drawing performance; and 3) comparable age range of the subjects.[12]

It is thus appropriate to examine the present research in an attempt to shed additional light on McCarthy's speculation that emotionally toned pretest activities may have contributed toward the unexpectedly low reliability coefficient which appeared.

The present findings seem to lend support to her explanation, since the group with the neutral pretest activity exhibited the greatest consistency of drawing performance, the groups with similarly emotionally toned tasks showed less consistency, and the groups with differently emotionally toned tasks showed the least consistent performance (significantly more variable in fact than the neutral group). Group IV, which

---

12 McCarthy's study included a somewhat broader age range since she used Grade 3 and 4 pupils, whereas the present study used only Grade 4 pupils. This is not felt to be a serious difference, however, in the light of McNemar's statement that "If we are drawing a sample from a group which is restricted in range with regard to either or both variables, the correlation will be relatively low." In McNemar, Q., Psychological Statistics, New York, Wiley, 1949, p. 125.

duplicated the conditions of McCarthy's subjects, exhibited the lowest reliability coefficient of the five groups of this study.

Shanan, in his study of intraindividual variability of drawing performance, remarked that

> . . . variability is high under all conditions which might be interpreted as anxiety arousing. Response consistency rises or levels off under conditions which may be presumed less anxiety arousing.[13]

In the present study a similar tendency is seen, although the affective state would not be interpreted as anxiety, but rather the subjects were thought to be affectively involved with either a positively or negatively toned task. It is possible however that intraindividual response variability increases throughout the general area of affective involvement, of which anxiety states are one aspect.

---

13 Shanan, op. cit., p. 110

SUMMARY AND CONCLUSIONS

Regarding the all-important question of the reliability of a test, studies which have investigated the retest reliability of the Goodenough Test since its appearance in 1926 have served to indicate the degree to which subjects' drawing performances are consistent on two occasions, varying the conditions of administration such as length of the interval.

Evidence appeared to suggest the importance of certain subject variables, such as the affective state of the individual, upon performance in the drawing situation. Group mean scores tended to rise or fall depending on the emotional tone of the activity, and variability of performance appeared to increase, as reflected by lowered correlation coefficients.

This study attempts to verify and extend existing knowledge about the possible influence of affective state in children on the variability of drawing performance as measured by the D.A.M. Test.

A sample of 261 English speaking Grade 4 children was divided into five groups, each group being given varying combinations of positively, negatively and neutrally toned pretest written tasks followed by the D.A.M. on both occasions.

From the results of this experiment it appears that drawing performance is somewhat less consistent over two occasions when preceded by similarly emotionally toned tasks than when preceded by neutral tasks, and significantly less

consistent when preceded by differently emotionally toned tasks than when preceded by neutral tasks. (Mean scores, however, did not differ significantly, nor did they exhibit predictable directional changes.) Thus if the activity prior to the D.A.M. Test is emotionally toned, the variability of the subject's performance may be significantly greater than it would be in circumstances where the activity prior to drawing is not emotionally toned.

In a clinical situation, then, where the D.A.M. Test is commonly introduced early as a "buffer" to help put the child at ease by involving him in a task which is familiar to him and cannot be "failed", knowledge of any activity prior to testing which might have evoked an affective reaction in the subject might be of help in evaluating the importance to be given to the D.A.M. within the total test battery.

Future research might be directed toward increasing the intensity of the pretest affective stimulation within a similar design to explore the critical factors involved in predictable changes of group mean scores.

Further investigation of variability of performance might lead in the direction of measuring the effect on intra-individual response variability of pretest activity whose nature varied to include many types of affective state.

BIBLIOGRAPHY

Goodenough, Florence L., The Measurement of Intelligence by Drawings, New York, World Book Company, 1926, xiv-177p.

This slim volume presents the Goodenough Draw A Man Test (D.A.M.), a highly useful tool which yields an estimate of a child's intellectual maturity from an easily obtained sample of drawing performance. On the basis of an historical survey the rationale of the test is developed.

Standardization was carried out on a large sample of children between the ages of six and twelve years, representing a wide range of cultural backgrounds and nationalities. The clearly outlined instructions for administration, scoring and evaluation of the test help to ensure a lasting place in the field of psychological measurement for this instrument.

Harris, D.B., Children's Drawings as Measures of Intellectual Maturity, New York, Harcourt, Brace & World, 1963, xvi-367p.

The author presents a revision and restandardization of the Goodenough Test, as well as an alternate form. Scoring points are clarified and increased in number.

There is detailed consideration of the psychology of drawing as a motor and cognitive act, as a projective device, and a discussion of the principal theories of children's drawings. The main value lies in the well organized presentation of an enormous body of research concerning human figure drawings.

McCarthy, Dorothea, "A study of the reliability of the Goodenough Drawing Test of Intelligence", in Journal of Psychology, Vol. 18, Second Half, issue of October 1944, p. 201-216.

This is a well controlled, clearly presented study which successfully achieved its main intention of investigating inter- and intra-scorer reliability. Unexpectedly low retest reliability coefficients led to speculation which suggested the present study.

Reichenberg-Hackett, W., "Changes in Goodenough drawings after a gratifying experience", in American Journal of Orthopsychiatry, Vol. 23, No. 3, issue of July 1953, p. 501-517.

The author investigates drawing performance following conditions of gratification, reward and neutral activity. Comparisons are made between scores of these groups and scores of a control group. Unfortunately the latter does not adequately fit the definition of a control group, thus throwing these comparisons in question. However the value of this research

lies in the author's valid demonstration that drawing scores tend to increase following a gratifying experience.

Smith, F.O., "What the Goodenough Intelligence Test measures", in Psychological Bulletin, Vol. 54, No. 9, issue of November 1957, p. 760-761.

This study serves to establish the level of retest reliability of the D.A.M. Test throughout its entire age range with a short interval between administrations, and to indicate that what is measured here is the more specific factor of mental maturity rather than the broader concept of general intelligence.

Shanan, J., "Intraindividual response variability in figure drawing tasks", Journal of Projective Techniques, Vol. 26, No. 1, issue of March 1962, p. 105-111.

The main focus of this study is on testing Fiske's hypothesis that greater intensity of external stimulation reduces response variability. Support was found for this hypothesis, and the author also indicated that variability increased under anxiety arousing conditions. This latter finding is of interest in connection with the present study.

Yepsen, L.N., "The reliability of the Goodenough Drawing Test with feeble-minded subjects", in Journal of Educational Psychology, Vol. 20, No. 6, issue of September 1929, p. 448-451.

Using a sample of feebleminded boys, the author examines the reliability of the D.A.M. Test and finds that consistency of the subjects' performances over three administrations with four day intervals is very high. The importance of this study lies in its demonstration of the usefulness of the D.A.M. Test for a sub-group which was not included in the standardization.

APPENDIX 1

SAMPLE COMPOSITION FORMS

APPENDIX 1

NAME _____ BOY___ GIRL___ I

BIRTHDATE _____ ___ ___ AGE _____ years old
Month Day Year

TASK A

Directions: You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.

"THE BEST THING THAT EVER HAPPENED TO ME"

APPENDIX 1

NAME _____ BOY___ GIRL___ I

BIRTHDATE _____ ____ _____ AGE _____ years old
Month Day Year

Alternate form of TASK A

Directions: You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.

"THE HAPPIEST DAY I CAN REMEMBER"

APPENDIX 1

NAME _____ BOY___ GIRL___

BIRTHDATE _____ ____ _____ AGE _____ years old
Month Day Year

TASK B

Directions: You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.

"THE WORST THING THAT EVER HAPPENED TO ME"

APPENDIX 1

NAME _____ BOY____ GIRL____ II

BIRTHDATE _____ _____ _____ AGE _____ years old
Month Day Year

Alternate Form of TASK B

Directions: You will have 15 minutes to write a one-page composition. If you need more space, write on the back of the paper. Go ahead and begin.

"THE SADDEST DAY I CAN REMEMBER"

NAME _____ **APPENDIX 1** BOY____ GIRL____ III

BIRTHDATE _____ _____ _____ AGE _____ years old
Month Day Year

TASK C

Directions: You will have 15 minutes to write all the words you can think of beginning with the letters shown. Go ahead and begin.

1. How many words can you think of beginning with the letters B or R ?

2. How many words can you think of beginning with the letters D or S ?

APPENDIX 1

NAME _____ BOY___ GIRL___ III

BIRTHDATE _____ _____ _____ AGE _____ years old
Month Day Year

Alternate form of TASK C

Directions: You will have 15 minutes to write all the words you can think of beginning with the letters shown. Go ahead and begin.

1. How many words can you think of beginning with the letters C or P ?

2. How many words can you think of beginning with the letters M or T ?

APPENDIX 2

PRESENTATION OF RAW SCORE DATA

APPENDIX 2

PRESENTATION OF RAW SCORE DATA

A preliminary analysis of the data was carried out based on the total point raw scores of the drawings.

Table V shows the mean scores and standard deviations for each group on test and retest.

Table VI presents Pearson *r* and Fisher *z* values for the five groups of subjects.

Table VII shows the critical ratio values resulting from testing the significance of differences among Fisher *z* values.

It was felt that the data would be more properly presented in the main body of this paper in terms of IQ values since 1) this is the usual method of presentation of data involving the D.A.M. Test; the results are therefore more easily compared with results of other studies, and 2) about twenty percent of the sample gained one month in terms of C.A. during the interval between test and retest. An adjustment in C.A. was made prior to calculation of retest IQ values for these individuals.

Comparison of correlation values of the raw score and IQ score computations shows that the total range of the Pearson *r*'s is considerably attenuated in the case of the IQ values, with the relationship between test and retest scores increasing

Table V.-

Means and Standard Deviations for First and Second Administration of the D.A.M. to Five Groups of Subjects[a]

Group	N	M_1	s_1	M_2	s_2
I	51	27.3	6.74	27.2	6.02
II	56	29.0	6.84	28.2	6.36
III	51	27.1	6.16	27.6	5.76
IV	54	27.8	5.73	28.1	6.17
V	49	28.3	5.54	29.0	5.68

a Calculations based on raw scores

Table VI.-

Test-retest Reliability Coefficients (Pearson *r*) and Fisher Transformation Values (*z*) for Five Subject Groups on the D.A.M. Test

Group	N	Pearson *r*	Fisher *z*
I	51	.733	.95
II	56	.804	1.10
III	51	.869	1.35
IV	54	.653	.74
V	49	.741	.95

Table VII.-

Critical Ratio Values used to Test Significance of Differences among Fisher z's

Group	I z=.95	II z=1.10	III z=1.33	IV z=.74
II z=1.10	.761			
III z=1.33	1.853[a]	1.107		
IV z=.74	1.05	1.855[a]	2.95[b]	
V z=.95	0.00	.75	1.002[a]	1.059

a p =.06
b Significant at p < .01

for Groups I, IV and V. Thus with the smaller difference between the highest and lowest correlations computed with I.Q. values, the difference between Groups III and IV is significant only at the p=.05, rather than $p < .01$ level, and the near-significant differences between Groups I and III; II and IV; and III and V found in the raw score calculations are lost when IQ values are used.

Sex Differences

Goodenough remarked that girls tend to score higher than boys on the D.A.M. scale.[1] Inspection of the data revealed that this held true in the present study, and in two groups where large differences were observed, 't' tests were used to establish the level of significance of the differences (data of the first administration).

The standard error of the difference was computed using the formula

$$\sigma_{Diff} = \sqrt{\frac{S_1^2}{N_1 - 1} + \frac{S_2^2}{N_2 - 1}}$$

Degrees of freedom varied between forty-seven and fifty-four, and the values for t(p=.05) and t(p=.01) were read from the 't' table.

---

1 Goodenough, Florence L., The Measurement of Intelligence by Drawings, New York, World Book Company, 1926, p. 30.

In Groups II and III, the difference between the boys' and girls' scores was significant at $p < .01$; however no further analysis was done to explore the implications of this finding.

APPENDIX 3

ABSTRACT OF
Emotional Tone of Pretest Activity
and Retest Reliability of the Goodenough
Draw A Man Test

APPENDIX 5

ABSTRACT OF

Emotional Tone of Pretest Activity and Retest Reliability of the Goodenough Draw A Man Test

This study investigated the possible effect of emotionally toned pretest activity on drawing performance of children, as measured by scores on the Goodenough Draw A Man Test (D.A.M.).

Previous studies suggested that emotionally toned pretest activity lowered the consistency of performance across two administrations of the D.A.M. Test, reflected by a lowered reliability coefficient. There was also evidence to suggest that drawing scores might increase following a gratifying experience and decrease under conditions of pleasant followed by unpleasant emotionally toned tasks.

The sample was composed of 261 children in ten Grade 4 classrooms in the English speaking Separate Schools of Ottawa. They were tested on two occasions, with an interval of one week between administrations. There were five subject groups, each being approximately equally represented in every classroom. The five groups were given differing combinations of positively, negatively and neutrally toned pretest activity, immediately followed by administration of the D.A.M. Test on each occasion. The pretest activities were written tasks, two of which were in composition form, the third being a word-listing task.

Scorer self-consistency was .92[illegible] (Pearson r) based on total point raw scores, and 89.62 in terms of mean percentage of agreement, based on a point-by-point analysis of fifty drawings. The results of the main analysis suggest that consistency of performance as reflected in retest reliability falls off from the most consistent performance of the neutral group to the less consistent performance of the groups having the same conditions of affective involvement on two occasions, to the least consistent performance of the groups experiencing change of affective involvement on two occasions. The combined reliability coefficient of the latter groups was significantly lower than that of the neutral group. Mean scores fluctuated in a random fashion.