R E P O R T   R E S U M E S

ED 016 340   40   EC 001 389

THE EFFECT OF AGE AS A VARIABLE ON THE SCORES OF THE HARRIS-GOODENOUGH DRAWING TEST OF EDUCABLE RETARDATES.

BY- LEVY, IRWIN S.
NORTH CAROLINA UNIV., CHAPEL HILL

REPORT NUMBER BR-6-8739   PUB DATE   AUG 67
GRANT   OEG-2-6-068739-1751
EDRS PRICE   MF-$0.50   HC-$2.96   72P.

DESCRIPTORS- *EXCEPTIONAL CHILD RESEARCH, *TESTS, *MENTALLY HANDICAPPED, ADOLESCENTS, EDUCABLE MENTALLY HANDICAPPED, TEST RELIABILITY, INTELLIGENCE TESTS, GROUP INTELLIGENCE TESTS, AGE DIFFERENCES, AGE, STANDARIZED TESTS,

IN ORDER TO DETERMINE THE RELIABILITY OF PERFORMANCE OF RETARDED ADOLESCENTS ON THE HARRIS REVISION OF THE GOODENOUGH DRAW-A-MAN TEST (DAM) AND WHETHER THE DECLINE IN PERFORMANCE WHICH OCCURS IN NORMAL ADOLESCENTS AT THE MID-TEENS ALSO OCCURS WITH RETARDED ADOLESCENTS, 213 MALE AND 130 FEMALE SUBJECTS, AGED 11-20 YEARS AND WITH IQ'S OF 56-72, IN INTERMEDIATE AND SECONDARY CLASSES FOR THE EDUCABLE MENTALLY HANDICAPPED (EMH) IN NORTH CAROLINA WERE TESTED. THE DAM WAS ADMINISTERED IN GROUP FORM TO ALL THE SUBJECTS IN THEIR OWN CLASSROOMS. IT WAS READMINISTERED AFTER 7 MONTHS. OVERALL MEAN CHANGE FOR THE 343 SUBJECTS BETWEEN TEST AND RETEST WAS SIGNIFICANT (P IS LESS THAN .05). ANALYSIS OF VARIANCE PRODUCED SIGNIFICANT F-RATIOS SHOWING THAT STANDARD DEVIATIONS OF THE CHANGE DIFFERED AT VARIOUS CHRONOLOGICAL AGE GROUPS FOR THE MALES. RESULTS INDICATED THAT THE TEST-RETEST RELIABILITY WAS SIGNIFICANT (P IS LESS THAN .01). THE TEST IS USEFUL WITH EMH FEMALES TO AGE 16 AND WITH EMH MALES TO AGE 20 YEARS. ALTHOUGH THE MAXIMUM CHRONOLOGICAL AGE DIVISOR OF 15 WAS ESTABLISHED BY HARRIS, THE INTRA-SCORER RELIABILITY COEFFICIENT AFTER 6 WEEKS WAS .99. IN CONCLUSION, THE DAM TEST AS A MEASURE OF CONCRETE CONCEPT FORMATION SEEMS TO BE A RELIABLE INSTRUMENT FOR GAINING INFORMATION ABOUT MILDLY MENTALLY HANDICAPPED ADOLESCENTS. TWENTY-FIVE REFERENCES AND 19 TABLES ARE INCLUDED. (DT)

FINAL REPORT

Project No. 6-8739
Grant No. OEG—2-6-068739-1751

The Effect of Age As A Variable On the Scores of the Harris-Goodenough Drawing Test of Educable Retardates

August 1967

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

The Effect of Age as a Variable on the Scores of the Harris-Goodenough Drawing Test of Educable Retardates

Project No. 6-8739
Grant No. OEG-2-6 068739-1751

Irwin S. Levy

August 1967

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

The University of North Carolina

Chapel Hill, North Carolina

ACKNOWLEDGMENTS

The writer wishes to thank those individuals whose contributions were of great importance to the course of this study.

To June, my wife, for her contributions as a research assistant, and as a source of inspiration during the process of this study, goes my everlasting devotion.

My chairman and consultant, Dr. Nancy M. Robinson, deserves more than my hearfelt appreciation for her genuine interest, keen insights, and continuous patience and support during the entire study.

Sincere gratitude goes to the other members of my graduate committee, Doctors Robert M. Anderson, Guy B. Johnson, Neill A. Rosser, and Luther R. Taff, whose supportive assistance was evident from the beginning of the study.

The directors of the Special Education Departments in the city school systems of Greensboro and Durham, North Carolina, were most helpful in providing subjects and confidential data for the study.

To professors Dale B. Harris and George S. Baroff go my sincere thanks for their interest in and support of the research undertaken.

The Division of Handicapped Children and Youth, Office of Education, U.S. Department of Health, Education, and Welfare deserves my deep appreciation and gratitude for its support in awarding me a research grant under number OEG-2-6 068739-1751, which gave additional impetus to the study.

TABLE OF CONTENTS

		Page
ACKNOWLEDGMENTS		ii
LIST OF TABLES		vi
CHAPTER		
I.	INTRODUCTION	1
	The Problem	2
	Objectives	4
II.	RESEARCH PROCEDURES	6
	Statement of Objectives	6
	The Sample	7
	The Instrument, Administration, and Scoring Procedures	9
	Methods of Analysis	10
III.	RESEARCH FINDINGS	12
	Intra-scorer Reliability	12
	Equivalence of Groups on Intelligence Quotients	12
	Relationship of the Two Instruments	19
	Test-Retest Reliability	19
	Alternate-Form Reliability	22
	Differences between Initial and Retest Standard Score IQs on the Drawing Test	26

TABLE OF CONTENTS--Continued

		Page
	Summary of Findings	37
IV.	DISCUSSION, CONCLUSIONS, AND IMPLICATIONS	39
	Discussion	39
	Relationship between the 1960 Stanford-Binet and the Harris-Goodenough Drawing Test (Full Scale)	40
	Test-Retest Reliability	44
	Alternate-Form Reliability	45
	Differences between Initial Test and Retest Standard Scores	46
	Conclusions	48
	Implications	49
V.	SUMMARY OF THE STUDY	51
	The Problem	51
	Review of the Literature	52
	Objectives	53
	Procedures	54
	Results	56
	Implications for Future Research	57
APPENDIX A		58
APPENDIX B		61
REFERENCES		63

LIST OF TABLES

TABLE Page

1. The Sample 8

2. Means and Standard Deviations of Stanford-Binet IQs by Age 13

3. Means and Standard Deviations of Stanford-Binet IQs by Age and Sex 14

4. Analysis of Variance of Stanford-Binet IQs 15

5. Initial Full Scale DAM IQs and Standard Deviations: Standard Scores and Interpolated Scores by Age . . 17

6. Initial Full Scale DAM IQ Means and Standard Deviations by Age and Sex 18

7. Analysis of Variance of Initial Full Scale DAM IQs 20

8. Correlations between Stanford-Binet and Initial Full Scale DAM IQs by Sex and Age 21

9. Test-Retest Reliability by Age 23

10. Test-Retest Reliability by Age and Sex . . . 24

11. Alternate-Form Reliability by Age and Sex 25

12. Differences between Initial Test and Retest Standard Scores (IQs) for the Man Scale and the Woman Scale by Age . . . 27

13. Differences between Initial Test and Retest Standard Scores (IQs) for the Man Scale by Age and Sex 28

14. Differences between Initial Test and Retest Standard Scores (IQs) for the Woman Scale by Age and Sex 30

LIST OF TABLES--Continued

TABLE Page

15. Differences between Initial Test and Retest Standard Scores (IQs) for Full Scale by Age 31

16. Differences between Initial Test and Retest Standard Scores (IQs) for Full Scale by Age and Sex 33

17. Analysis of Variance of the Differences in Full Scale Test-Retest Scores by Age: CAs 17-20 Combined 34

18. Analysis of Variance of Difference in Full Scale Test-Retest Scores of Male Ss: CAs 17-20 Combined 35

19. Analysis of Variance of Difference in Full Scale Test-Retest Scores of Female Ss: CAs 16-20 Combined 36

CHAPTER I

INTRODUCTION

Although published more than forty years ago, the Goodenough Draw-a-Man Test (DAM) remains a popular and widely used method for the assessment of intelligence (Sundberg, 1961). Of the many studies utilizing the Goodenough scale, no population has been used more widely than that of the mentally retarded (Kennedy & Lindner, 1964, p. 36). Its continued use has received recent impetus from the publication of a modern revision of the instrument.

The original Goodenough test consisted of the drawing of a man, scoring of which was standardized on a fairly representative population of children, ages 3 to 13.

Dale B. Harris (1963) revised and restandardized the 1926 Goodenough DAM test, making the following changes: (a) extension of the chronological age (CA) range of the norms from 13-0 to 15-11; (b) addition and standardization of the drawing of a woman as an alternate form of the drawing test; (c) increase in the number of raw score points from 51 to 73 on the Man scale and the addition of 71 points on the Woman scale; (d) alteration of the concept of mental age (MA) to a percentile rank and conversion of the ratio IQ to a deviation IQ with a mean of 100 and a standard deviation of 15; and (e) the inclusion of a self-scale drawing which has not yet been standardized, as a further measure of mental maturity.

As of 1967, the reliability of the performance of retarded subjects has not been examined on the revised edition of the DAM test.

The Problem

The reliability of the performance of retarded children and adolescents on the revised version of the DAM has not been explored, and it is not clear whether the decline in performance which occurs in normal adolescents in mid-teens also typifies the performance of retarded individuals (Robinson & Robinson, 1965, p. 434). The Goodenough test scores cease to show increments soon after Bayley's (1956) "manipulating symbols" period of mental development terminates and during Piaget's (1953) shift from concrete operations to formal operations. This suggests that the drawing test evaluates primarily the ability to form concrete concepts (Harris, 1963).

This cessation in increments in score is considered by Harris (1963) as being dependent on three possible phenomena of adolescence: (a) the increasing psychological and motivational conflicts, particularly over bodily changes and sex; (b) the preeminence of language in its increasing ability to delineate cognitive content or concepts; and (c) the child's increasing ability to judge his own drawing as a conceptualization and representation of a visual reality and his increasing self-criticism of technique. Support of the first position appears to be more negative than positive. The second view is founded partially in psychological and sociological evidence such as Buhler (1930) and Vinacke (1954). For the third position, persuasive psychological evidence indicates that the child grows self-critical because of his inability to reproduce photographic likeness in his art work and may abandon drawing altogether as a mode of communication (Harris, 1963).

Unless the child can master the techniques that are necessary for the achievement of realistic drawings at this stage, the drawing of human figures as a measure of mental maturity ceases to be valid at this point (Harris, 1963).

The very young child experiences more or less directly the primary qualities of concrete objects. Undoubtedly, the concept of a person as a concrete object undergoes an elaborate evolvement and differentiation with age. The child moves into Piaget's period of formal operations when his intellectual processes become sufficiently advanced and complex for him to conceptualize abstract, logical, and hypothetical relationships. His thinking and his visual-motor production become characterized by more complex and abstract qualities. Probably because it taps more concrete concepts, the drawing test at this time ceases to show increments and therefore ceases to be an index to the child's further growth in intellectual maturity (Harris, 1963, p. 7).

Mitchell (1959) gave some evidence that increase in test performance scores of mentally retarded subjects, however, continued throughout adolescence on the original Goodenough test. Evidence from tests of intelligence other than the DAM clearly indicates that mental age in mentally retarded persons as well as those with normal intelligence continues to increase well beyond thirteen years (Mitchell, 1959, p. 555). The age range of the Harris-Goodenough DAM test has been extended only from 13-0 to 15-11. If the DAM test score continues to increase after age 15, IQs of mentally retarded adolescents might be overestimated by an artificial restriction of their CA increment. A gradual and spurious increase in IQ from CA 15-11 to that point at which MA growth ceases would be noted.

Objectives

Although Goodenough confirmed that the DAM test ceased to show increments in scores of children of normal intelligence, by early adolescence, it might be possible to devise a special standardization above that CA for a retarded population. This study was undertaken to determine the most appropriate CA divisor to be used with subjects (Ss) whose MA renders them suitable candidates for the drawing test, but whose CA is greater than those in Harris' restandardization population. The following questions were asked in this investigation:

(1) Can the test be reliably scored?

(2) Are the IQs obtained on the Harris-Goodenough revision of the drawing test stable over a seven-month interval? Is there age variability in such restest stability?

(3) At what age, if any, in this range, do test scores obtained by educable mentally retarded adolescents cease to increase on a retest after a seven-month interval?

(4) What is the most appropriate CA divisor to employ in calculating the IQs of such Ss above the presently recommended CA at 15?

Overview of the Study

Chapter II discusses the research, procedures, subjects, and methods of analysis.

Chapter III reports the research findings.

Chapter IV discusses and summarizes the research findings and offers implications and suggestions for further research.

Chapter V summarizes the study.

CHAPTER II

RESEARCH PROCEDURES

Statement of Objectives

Although Goodenough (1926) confirmed that the drawing test ceases to show increments by age in early adolescence for children of normal intelligence, this finding has not been replicated in a retarded population on the Harris (1963) revision of the DAM test. This study was initiated to determine the most appropriate CA divisor to use with retarded children whose MAs render them suitable candidates for the drawing test, but whose CAs are greater than those of the Ss in Harris' restandardization population.

The following questions were asked in this experiment:

(1) Can the test be reliably scored?

(2) Are the IQs obtained on the Harris-Goodenough revision of the DAM test stable over a seven-month period? Is there variability by age and sex in such retest stability?

(3) At what age, if any in this range, do the test scores obtained by educable mentally retarded adolescents

cease to increase on retest after a seven-month interval?

(4) What is the most appropriate CA to employ in calculating the IQs of such subjects above the presently recommended maximum CA of 15-11?

The Sample

The sample consisted of 572 Ss enrolled in intermediate and secondary special classes for educable mental retardates in the public school systems of Greensboro and Durham, North Carolina. At the time of the retest only 343 Ss of the original group could still be found in the public schools. These 343 Ss comprised the sample used in the study. There were 213 male Ss and 130 female Ss ranging in age from 11-0 to 20-6 at the time of the first test. Table 1 gives the number of each sex per age group.

Placement in special classes is determined by an intelligence test, either the Stanford-Binet or Wechsler Intelligence Scale for Children (WISC). A child with an IQ between 50 and 75 is eligible for special class placement. These individual tests are administered by state certified psychometrists or professional

TABLE 1

THE SAMPLE

CA	Male	Female	Total
11	30	12	42
12	34	25	59
13	38	28	66
14	47	25	72
15	27	16	43
16	15	8	23
17	4	7	11
18	10	5	15
19	3	3	6
20	5	1	6
T	213	130	343

psychologists and are repeated at approximately three-year intervals.

All Ss used in this sample were free from known sensory or physical handicaps. Classification of educable mental retardation was made without regard to categories such as neurological impairment, central nervous system disorder, or cultural-familial diagnosis. No attempt was made to classify students on the basis of race or socioeconomic variables.

Each S's IQ was used in conjunction with his actual CA at the time of the first DAM test administration in order to arrive at a current MA. These MAs were derived from the IQ tables in the 1960 revision of the Stanford-Binet (Terman & Merrill, 1960).

The Instrument, Administration, and Scoring Procedures

The Harris-Goodenough Drawing Test was used in the study to evaluate the objectives proposed. The DAM test was administered in group form by the classroom teacher to all intermediate and secondary educable mentally retarded classes in the two school systems following the written instructions from the investigator (See Appendix B).

On each test administration, each subject was given two sheets of 8 1/2 x 11 inch plain white, bond paper with his name in the upper left hand corner. Plain paper was used rather than the suggested test booklet to reduce test anxiety. A man and a woman drawing were secured from each S on each administration. The retest was made following a seven-month interval under the same standard procedures. During this time, regular academic classwork was performed, and no attempt was made to train, coach, or influence the retest scores.

From the school records, additional data were secured, including each S's individual IQ score, test, and form; CA; and MA.

A single trained examiner (JL) unaware of the previous IQ of any S scored all the drawings so that any errors in scoring were presumably consistent. Coded numbers were used for Ss' names. Scoring was done by crediting each appropriate item of the pretest and posttest. Raw scores were not tabulated until all four drawings of each S were scored, and, subsequently, standard scores were not converted until all raw scores had been tabulated. This was done to maintain scorer objectivity and to reduce the influence of the scores of the first administration.

Methods of Analysis

The raw data were converted to standard scores using the tables established by Harris (1963); however, since the test was administered with only a seven-month interval instead of a whole year, standard scores for Ss in CA groups 11-14 were interpolated to reflect the proportional change in CA growth.

Since the standard scores proposed by Harris center at the mid-point of each CA year, and the change in them was not linear, a conversion table based on difference in age in months above and below the mid-point was constructed to indicate the true change in standard score for individual Ss. Since the ceiling of the standard scores is reached at CA 15-11, only those Ss who had not attained that age by the retest had their test scores interpolated. The interpolated values were rounded to the nearest whole number with the exception of those values which resulted in .5. Odd numbered standard scores were increased to the next even number, and even numbered standard scores dropped the fraction to avoid systematic influence on the scores.

Analysis of variance was computed on initial DAM and Stanford-Binet IQs.

Correlations were computed for test-retest reliability, alternate-form reliability, intra-scorer reliability, and test validity on the mentally retarded population using the Pearson product-moment formula. Comparison of correlations utilized Fisher's r to z transformation (Edwards, 1964).

CHAPTER III

RESEARCH FINDINGS

Computations were made by the Computation Center of the University of North Carolina at Chapel Hill and by the investigator.

Intra-scorer Reliability

Six weeks after the original scoring, the investigator selected at random 25 drawings of a woman and 25 drawings of a man to be rescored by the research assistant. The correlations between first and second scorings were .99 for each set of test scores.

Equivalance of Groups on Intelligence Quotients

Tables 2 and 3 report the mean IQs and standard deviations of the Ss' IQs on the 1960 Stanford-Binet Scale, grouped by age and sex respectively.

Analysis of variance of the Stanford-Binet IQs according to age groups indicated an F-ratio of 2.237 ($p < .05$). Table 4 contains the results of this analysis of variance.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF

STANFORD-BINET IQs BY AGE

CA	N	Mean	Standard Deviation
11	42	64.9	9.2
12	59	64.4	6.7
13	66	65.7	6.3
14	72	65.9	6.9
15	43	61.9	8.4
16	23	63.3	8.1
17	11	60.9	8.4
18	15	66.1	6.3
19	6	70.7	9.6
20	6	63.2	4.4
TOTAL	343	64.7	7.5

TABLE 3

MEANS AND STANDARD DEVIATIONS OF

STANFORD-BINET IQs BY AGE AND SEX

	Male			Female		
CA	N	Mean	Standard Deviation	N	Mean	Standard Deviation
11	30	66.0	9.1	12	62.3	9.0
12	34	66.0	6.2	25	62.4	7.0
13	38	67.1	5.8	28	63.9	6.7
14	47	66.1	7.6	25	65.6	5.4
15	27	61.0	7.4	16	63.4	10.0
16	15	61.9	7.2	8	66.0	9.4
17	4	56.3	5.9	7	63.6	8.8
18	10	65.9	7.7	5	66.4	2.9
19	3	72.0	11.5	3	69.3	9.5
20	5	63.6	4.8	1	61.0	---
TOTAL	213	65.1	7.5	130	64.0	7.4

TABLE 4

ANALYSIS OF VARIANCE OF STANFORD-BINET IQs

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	1,077	9	119.7	2.237*
Within groups	17,802	333	53.5	
TOTAL	18,879	342		

*p < .05

Using the formula $\frac{(\bar{X}_1-\bar{X}_2)^2}{s_w(n_1+n_2)/(n_1)(n_2)}$

(Ferguson, 1966), Scheffe's test of multiple comparisons revealed no significant differences between the means of any of the CA groups. Comparison of extreme variances within the age groups indicated a significant F-ratio ($p < .01$) between CA groups 13 and 19 for male Ss on Stanford-Binet IQs, as indicated in Table 3. The F-ratios of the variances were not significant for other CA groups on Stanford-Binet IQs. Since no pair of means were significantly different according to Scheffe's test, the CA groups appear to come from equivalent populations. It is possible that the significant difference in variance may be due to chance.

Tables 5 and 6 report the initial Full Scale DAM IQs by age and sex respectively. Since there was only a 7 month test-retest interval, and because Harris' standard scores are based on a 12 month interval, interpolated IQs were computed to reflect the proportional change in CA growth of those Ss who had not reached CA 15 at the time of the retest. By means of visual inspection no significant differences between the two sets of scores were seen; therefore, further analyses were made only on standard scores.

The standard deviations of the total group as well as for both male Ss and female Ss is somewhat lower than

TABLE 5

INITIAL FULL SCALE DAM IQs AND STANDARD DEVIATIONS:

STANDARD SCORES AND INTERPOLATED SCORES BY AGE

		Standard Scores		Interpolated Scores*	
CA	N	Mean	Sd	Mean	Sd
11	42	81.8	11.7	81.9	11.7
12	59	79.0	11.3	79.0	11.2
13	66	78.0	11.9	78.2	11.9
14	72	77.4	13.1	77.7	13.1
15	43	74.8	14.1		
16	23	80.1	15.7		
17	11	81.5	9.0		
18	15	76.5	12.3		
19	6	87.7	18.4		
20	6	73.3	11.5		
11-14	239	78.7	12.1		
15-20	104	77.6	14.1		
TOTAL	343	78.4	12.7	78.9	12.1

*No significant differences between standard scores and interpolated scores were revealed; hence, further computations and analyses were based on standard scores.

TABLE 6

INITIAL FULL SCALE DAM IQ MEANS AND STANDARD DEVIATIONS BY AGE AND SEX

	Male			Female		
CA	N	Mean	Standard Deviation	N	Mean	Standard Deviation
11	30	85.4	10.4	12	72.7	10.0
12	34	81.8	11.4	25	75.0	10.0
13	38	82.3	12.1	28	72.1	8.7
14	47	78.6	12.5	25	75.0	14.0
15	27	78.1	15.1	16	69.3	10.6
16	15	80.6	13.4	8	79.3	20.3
17	4	80.0	8.5	7	82.4	9.8
18	10	76.9	11.9	5	75.6	14.4
19	3	85.7	19.2	3	89.7	21.6
20	5	74.0	12.7	1	73.0	----
TOTAL	213	80.8	12.5	130	74.4	12.2

the standard deviation of 15 established by Harris (1963). This difference is to be expected, because the subjects were confined to a narrow IQ range.

Table 7 shows an F-ratio of 1.541 on analysis of variance of initial Full Scale IQs which did not reach statistical significance at the .05 level. The age groups, therefore, did not differ in initial Full Scale IQ on the DAM to a greater degree than expected by chance.

Relationship of the Two Instruments

The analysis of the relationship of the initial Full Scale DAM to the 1960 Stanford-Binet revealed a correlation of .27 for the total group as indicated in Table 8. Correlations between the two tests for male Ss was .25 and for female Ss was .27. The range of correlations by age groups for the Full Scale DAM and Stanford-Binet was from -.13 to .76. The range for male Ss was from .00 to .62 while the range for female Ss was from -.42 to .99. The correlations which reached levels of significance ($p < .01$) when grouped by age are of modest size. While only the significant $r = .82$ for CA 11 female Ss is substantial ($p < .01$), it was computed on a small number of cases.

Test-Retest Reliability

Coefficients of correlation indicated significant and substantial reliability on a test-retest basis for the

TABLE 7

ANALYSIS OF VARIANCE OF INITIAL

FULL SCALE DAM IQs

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	2,229	9	247.7	1.541ns
Within groups	53,311	333	160.1	
TOTAL	55,540	342		

TABLE 8

CORRELATIONS BETWEEN STANFORD-BINET AND INITIAL FULL SCALE DAM IQs BY AGE AND SEX

	Male		Female		Total	
CA	N	r	N	r	N	r
11	30	.32	12	.82***	42	.47***
12	34	.04	25	.40*	59	.25
13	38	.06	28	.11	66	.17
14	47	.37**	25	.26	72	.33***
15	27	.28	16	.36	43	.23
16	15	.16	8	-.42	23	-.13
17	4	.00	7	.49	11	.39
18	10	.15	5	-.22	15	.07
19	3	.62	3	.99	6	.76
20	5	.58	1	---	6	.20
T	213	.25***	130	.27***	343	.27***

*$p < .05$
**$p < .02$
***$p < .01$

drawing test when analyzed by both age and sex.

Table 9 shows the range of the retest coefficient of stability on the Full Scale to be from .62 to .90 with the total group r = .81. This is significantly higher ($p < .05$) than the reliability of .73 reported for the Man Scale alone. There was no significant difference between the Full Scale (.81) and the Woman Scale (.78) for the total group on the retest correlations.

When analyzed by sex differences, the male Ss' test-retest correlation of .69 on the Man Scale and their correlation of .80 for the Full Scale differed significantly ($p < .01$). Male Ss demonstrated a significant difference ($p < .05$) between test-retest of the Man Scale (.69) and test-retest of the Woman Scale (.77). Female Ss demonstrated no significant test-retest differences between the Man Scale (.75) and the Woman Scale (.74) or the Full Scale (.80). This finding indicates that the male Ss had significantly higher retest reliability for the Full Scale than the Man Scale while female Ss showed no significant difference among the three scales (see Table 10).

Alternate-Form Reliability

Separate correlation coefficients between the Man and Woman scales for the initial test and the retest were calculated (see Table 11).

When analyzed by total groups, the retest alternate-form reliability of .81 was significantly higher ($p < .05$)

TABLE 9

TEST-RETEST RELIABILITY BY AGE

		Man Scale	Woman Scale	Full Scale
CA	N	r	r	r
11	42	.75***	.89***	.89***
12	59	.68***	.74***	.76***
13	66	.84***	.76***	.90***
14	72	.75***	.81***	.82***
15	43	.71***	.79***	.80***
16	23	.76***	.85***	.86***
17	11	.83***	.63*	.77***
18	15	.27	.65***	.62**
19	6	.80	.82*	.81*
20	6	.67	.73	.77
TOTAL	343	.73***	.78***	.81***

*p < .05
**p < .02
***p < .01

TABLE 10

TEST-RETEST RELIABILITY BY AGE AND SEX

		Male				Female		
CA	N	Man Scale	Woman Scale	Full Scale	N	Man Scale	Woman Scale	Full Scale
11	30	.73***	.93***	.90***	12	.60*	.67**	.72***
12	34	.63***	.78***	.77***	25	.76***	.55***	.69***
13	38	.84***	.69***	.88****	28	.91***	.79***	.86***
14	47	.74***	.80***	.82***	25	.75***	.82***	.81***
15	27	.69***	.80***	.82***	16	.71***	.63***	.69***
16	15	.69***	.73***	.74***	8	.89***	.95***	.97***
17	4	.81***	.44	.67	7	.87**	.73	.80*
18	10	.14	.76**	.63*	5	.55	.49	.69
19	3	.46	.75	.63	3	.99	.98	.99
20	5	.95***	.75	.79	1	---	---	---
T	213	.69***	.77***	.80***	130	.75***	.74***	.80***

*p < .05
**p < .02
***p < .01

TABLE 11

ALTERNATE-FORM RELIABILITY BY AGE AND SEX

	Male			Female			Total		
CA	N	Initial Test	Retest	N	Initial Test	Retest	N	Initial Test	Retest
11	30	.78***	.86***	12	.60*	.77***	42	.79***	.88***
12	34	.66***	.72***	25	.76***	.83***	59	.69***	.70***
13	38	.63***	.81***	28	.60***	.77***	66	.68***	.84***
14	47	.77***	.82***	25	.82***	.80***	72	.78***	.81***
15	27	.91***	.85***	16	.69***	.75***	43	.86***	.85***
16	15	.92***	.65***	8	.88***	.86***	23	.90***	.74***
17	4	.84	.97*	7	.72	.74	11	.74***	.72***
18	10	.53	.76**	5	.57	.39	15	.54*	.67**
19	3	.97	.99	3	.99	.95	6	.97***	.90**
20	5	.79	.79	1	---	---	6	.78	.81*
T	213	.74***	.80***	130	.75***	.79***	343	.75***	.81***

*$p < .05$
**$p < .02$
***$p < .01$

than the initial test alternate-form reliability of .75. Correspondence between the two scales was, therefore, apparently greater during final testing.

When correlations were examined for each sex separately, alternate-form reliability for the male Ss on initial test was .74 and on retest was .80. The reliability coefficients for female Ss were .75 for the initial test and .79 for the retest. All of the reported correlations were significant at the .01 level.

Differences between Initial and Retest Standard Score IQs on the Drawing Test

Man Scale.--The total sample, on retest, showed an average gain of 2.7 points on the Man Scale as seen in Table 12. This difference between the means was statistically significant at the .05 level (t = 2.58).

Standard scores of all CA groups increased slightly on the Man Scale at the end of the 7-month interval except for the CA 18 group.

When analyzed separately by sex, the male Ss decreased in score only at the CA 18 level, while female Ss decreased in scores at CAs 16 and 18, as shown in Table 13.

The means of the amounts of change in standard scores, a 3.6 point increase for male Ss and a 1.2 increase for female Ss, are significantly different at the .05 level (t = 2.21). A significant difference in favor of the

TABLE 12

DIFFERENCES BETWEEN INITIAL TEST AND RETEST STANDARD SCORES (IQs) FOR THE MAN SCALE AND THE WOMAN SCALE BY AGE

		Man Scale				Woman Scale			
CA	N	Test	Retest	Difference	Sd of Difference	Test	Retest	Difference	Sd of Difference
11	42	82.0	83.5	1.5	8.8	81.0	84.0	3.0	6.6
12	59	77.8	80.5	2.7	9.3	79.6	81.4	1.8	9.3
13	66	78.4	82.1	3.7	7.6	77.0	78.4	1.4	9.4
14	72	76.7	79.1	2.4	9.9	77.5	77.5	0.1	9.4
15	43	75.1	79.8	4.7	12.4	74.1	80.1	6.0	11.4
16	23	78.1	80.1	2.0	12.1	81.5	81.0	-0.5	7.7
17	11	81.9	84.3	2.4	8.1	80.7	78.6	-2.1	8.5
18	15	79.6	75.7	-3.9	16.1	72.9	73.7	0.8	10.9
19	6	85.2	86.5	1.3	9.6	90.0	83.3	-6.7	12.3
20	6	77.0	86.5	9.5	10.3	69.2	80.0	10.8	8.0
TOTAL	343	78.2	80.9	2.7	10.1	78.0	79.7	1.7	9.6

TABLE 13

DIFFERENCES BETWEEN INITIAL TEST AND RETEST STANDARD SCORES (IQs) FOR THE MAN SCALE BY AGE AND SEX

	Male					Female				
	N	Initial Test	Retest	Difference	Sd of Difference	N	Initial Test	Retest	Difference	Sd of Difference
11	30	84.9	86.9	2.0	8.8	12	74.8	75.0	0.2	9.1
12	34	79.4	82.8	3.4	10.2	25	75.6	77.4	1.8	8.1
13	38	83.1	87.6	4.5	7.8	28	72.0	74.7	2.5	7.2
14	47	77.4	80.0	2.6	9.5	25	75.5	77.4	1.9	10.7
15	27	77.6	84.5	6.9	13.4	16	70.9	71.9	1.0	10.0
16	15	78.7	82.3	3.6	12.2	8	77.1	75.9	-1.2	12.0
17	4	80.3	85.5	5.2	8.7	7	82.9	83.6	0.7	7.8
18	10	78.9	76.5	-2.4	17.5	5	81.0	74.0	-7.0	14.1
19	3	82.3	84.7	2.4	14.8	3	88.0	88.3	0.3	2.5
20	5	78.2	88.8	10.6	11.2	1	71.0	75.0	4.0	----
TOTAL	213	80.1	83.7	3.6	10.6	130	75.1	76.3	1.2	9.1

male Ss is seen also between the means of the initial Man Scale ($t = 4.08$, $p < .01$) and between the mean standard scores of the retest ($t = 5.05$, $p < .01$).

Woman Scale.--The total sample, on the retest, showed an average gain of 1.7 points on the Woman Scale as indicated by Table 12.

Scores of the CA groups increased slightly on the Woman Scale except for the CA groups 16, 18 and 19. When analyzed separately by sex, male Ss showed slight increases in scores at CAs 11, 12, 14, 15, 16, and a large increase at CA 20. Female Ss increased slightly at CAs 11, 12, 15, 18, and 20. The male Ss demonstrated a mean gain of 2.3 points, while the female Ss showed an increase of 0.7 standard score points. The difference in the increase in scores between the sexes was not statistically significant. A significant difference in favor of the male Ss, however, was seen between the means of the initial Woman Scale standard scores ($t = 4.75$, $p < .01$) and the means of the retest on the Woman Scale ($t = 5.97$, $p < .01$).

Full Scale.--The total sample showed an average increase of 2.2 points on the Full Scale as seen in Table 15. This difference between the means of the pretest and posttest is statistically significant ($t = 2.18$, $p < .05$). Since the Full Scale is the average of the standard scores of both the Man Scale and the Woman Scale, the gain in points on the Full Scale is a reflection of the performance of both sex groups on

TABLE 14

DIFFERENCES BETWEEN INITIAL TEST AND RETEST STANDARD SCORES (IQs) FOR THE WOMAN SCALE BY AGE AND SEX

	Male					Female				
CA	N	Initial Test	Retest	Difference	Sd of Difference	N	Initial Test	Retest	Difference	Sd of Difference
11	30	85.4	87.9	2.5	4.8	12	70.0	74.2	4.2	9.9
12	34	83.7	85.7	2.0	8.3	25	73.9	75.5	1.6	10.6
13	38	81.0	84.1	3.1	11.0	28	71.7	70.7	-1.0	6.5
14	47	79.3	78.9	-0.4	9.4	25	73.9	74.9	1.0	9.5
15	27	78.3	85.9	7.6	11.4	16	67.0	70.4	3.4	11.4
16	15	81.9	82.1	0.2	8.4	8	80.8	79.1	-0.7	6.3
17	4	79.0	76.8	-2.2	8.5	7	81.7	79.7	-2.0	9.2
18	10	74.3	74.2	-0.1	9.8	5	70.2	72.6	2.4	14.0
19	3	88.7	86.3	-2.4	16.0	3	91.3	80.3	-11.0	7.9
20	5	69.4	81.4	12.0	8.3	1	68.0	73.0	5.0	----
TOTAL	213	80.9	83.2	2.3	9.6	130	73.3	74.0	0.7	9.5

TABLE 15

DIFFERENCES BETWEEN INITIAL TEST AND RETEST STANDARD SCORES (IQs) FOR THE FULL SCALE BY AGE

CA	N	Initial Test	Retest	Mean Difference	Sd of Difference
11	42	81.8	84.0	2.2	6.2
12	59	79.0	81.2	2.3	7.9
13	66	78.0	80.5	2.5	5.8
14	72	77.4	78.7	1.3	8.3
15	43	74.8	80.2	5.4	10.5
16	23	80.1	80.8	0.7	8.1
17	11	81.5	81.6	0.0	7.4
18	15	76.5	75.0	-1.5	10.4
19	6	87.7	85.2	-2.5	10.7
20	6	73.3	83.7	10.3	7.6
TOTAL	343	78.4	80.6	2.2	8.1

the individual scales. Table 15 indicates that score increments are shown through CA 16 although the difference earned by any CA group is slight.

When the differences in Full Scale retest scores were analyzed by each sex independently, some immediate differences were noticed. Female Ss ceased to gain increments at CA 16 while male Ss ceased to gain increments in standard scores at CA 18. However, there is a sudden and unexpected increase at CA 20 as shown in Table 16.

To test the significance of the contribution of age to the mean increase in retest scores, an analysis of variance was computed on the total group as well as on each sex group. Since there was a small number of Ss in each of the older CA groups, CAs 17-20 were combined for the total sample and male analyses, while CAs 16-20 were combined for the female sample. This procedure reduced the degrees of freedom between groups and permitted a more stable estimate of the variance within the older groups.

Table 17, which reports the data on the total sample, yields an F-ratio of 2.03 which did not reach significance at the .05 level. Table 18 shows an analysis of variance computed for the male Ss, which yielded an F-ratio of 2.52, significant beyond the .05 level.

The F-ratio of 15.45 computed on the female Ss was highly significant beyond the .01 level of confidence, as seen in Table 19.

TABLE 16

DIFFERENCES BETWEEN INITIAL TEST AND RETEST STANDARD SCORES (IQs) FOR THE

FULL SCALE BY AGE AND SEX

	Male					Female				
CA	N	Initial Test	Retest	Difference	Sd of Difference	N	Initial Test	Retest	Difference	Sd of Difference
11	30	85.4	87.6	2.2	5.7	12	72.7	74.8	2.2	7.7
12	34	81.8	84.5	2.6	7.5	25	75.0	72.9	1.8	8.5
13	38	82.3	86.1	3.8	6.3	28	72.1	72.9	0.8	4.7
14	47	78.6	79.8	1.2	8.0	25	75.0	76.5	1.5	8.8
15	27	78.1	85.5	7.3	10.7	16	69.3	71.4	2.1	9.5
16	15	80.6	82.5	1.9	9.2	8	79.3	77.6	-1.6	5.1
17	4	80.0	81.3	1.3	7.3	7	82.4	81.9	-0.7	8.0
18	10	76.9	75.6	-1.3	10.9	5	75.6	73.8	-1.8	10.5
19	3	85.7	86.0	0.3	15.6	3	89.7	84.3	-5.3	4.5
20	3	74.0	85.6	11.6	7.0	1	70.0	74.0	4.0	----
TOTAL	213	80.8	83.7	3.0	8.3		74.4	75.4	1.0	7.7

TABLE 17

ANALYSIS OF VARIANCE OF THE DIFFERENCE IN FULL SCALE TEST-RETEST SCORES BY AGE: CAs 17-20 COMBINED

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	667.26	6	111.21	
				2.03ns
Within groups	18,365.66	336	54.66	
TOTAL	19,032.92	342		

TABLE 18

ANALYSIS OF VARIANCE OF THE DIFFERENCE IN FULL SCALE TEST-RETEST SCORES OF MALE SUBJECTS CAs 17-20 COMBINED

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	837.33	6	139.6	
				2.52*
Within groups	11,398.10	206	55.3	
TOTAL	12,235.43	212		

*p < .05

TABLE 19

ANALYSIS OF VARIANCE OF THE DIFFERENCE IN FULL SCALE TEST-RETEST SCORES OF FEMALE SUBJECTS CAs 16-20 COMBINED

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	3,875.06	5	775.01	
				15.45***
Within groups	6,218.93	124	50.15	
TOTAL	10,093.99	129		

***p < .01.

The results of this analysis of variance reflect the fact that the older CA female Ss (16-20) had a decrement in retest scores which differed significantly from the younger CA female Ss (11-15), all of whom had increments in retest scores.

Summary of Findings

1. The self-scorer agreement after a six-week interval from the original scoring was .99 on both Man and Woman Scales.

2. Correlation coefficient between the 1960 Stanford-Binet Scale and the Harris-Goodenough Drawing Test was .27 for 343 Ss. Coefficients of .25 for 213 male Ss and .27 for 130 female Ss were reported.

3. Test-retest reliability for 343 Ss was .81 ($p < .01$) for the Full Scale, .73 ($p < .01$) for the Man Scale, and .78 ($p < .01$) for the Woman Scale.

4. Test-retest reliabilities for male Ss and female Ss were comparable, .80 ($p < .01$), on the Full Scale. Test-retest reliability for male Ss on the Man Scale was .69 and on the Woman Scale was .75. For female Ss, coefficients of test-retest reliability were .75 on the Man Scale and .74 on the Woman Scale. All coefficients were statistically significant ($p < .01$).

5. Alternate-form reliability for 343 Ss was .75 on the initial test and .81 on the retest. Male Ss demonstrated alternate-form reliability coefficients of .74

on the initial test and .80 on the retest, while female Ss obtained alternate-form reliability coefficients of .75 and .79 on the two test administrations respectively. All reached significance at the .01 level.

6. Difference in scores for the total sample, on retest of the Man Scale was 2.7 points. Male Ss increased their retest scores by an average of 3.6 points while female Ss had an average increase of 1.2 points.

7. Difference in scores for the total sample, on retest of the Woman Scale was 1.7 point increase for 343 Ss. Male Ss had an average increase of 2.3 points while female Ss gained 0.7 points, or almost no change in initial-retest scores. The male Ss' increase was not significantly higher on the Woman Scale.

8. The total sample showed an average increase of 2.2 points on the Full Scale. Male Ss had an average increase of 3.0 points and female Ss had an increase of 1.0 points on the Full Scale. Male Ss' increase in scores, although slight, continued through CA 18, while female Ss' scores ceased to increase at CA 16. Although male Ss decreased in score increments at CA 19, there was a large and sudden increase of scores at CA 20 (11.6 points).

CHAPTER IV

DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

Discussion

Intra-scorer reliability.--The intra-scorer reliability of .99 is slightly higher than the intra-scorer reliability of .94 reported by McCarthy (1944) using the original Goodenough Draw-a-Man Test with children of normal intelligence. Harris (1963), however, using two independent scorers for 150 drawings of a man and a woman, reported the very high correlations of .98 and .97 for the Man Scale and Woman Scale respectively. It thus appears that the new scoring standards produce highly reliable scores.

It was apparent that many of the rescored drawings, while representative of all levels of performance for the present sample, involved only items which required a minimum of difficult decision making by the scorer. Some of the more advanced items on the scales do involve more difficult judgments so that typical intra-scorer reliability of less highly trained examiners assessing productions by adolescents of normal or superior intelligence, might be somewhat lower.

Relationship between the 1960 Stanford-Binet and the Harris-Goodenough Drawing Test (Full Scale)

Harris (1963, p. 107) recommends combining the values of the Man and Woman Scales (Full Scale) to give a more reliable estimate of test achievement. The average thus obtained, he states, is a statistically more accurate estimate of the ability measured by the drawing test than that obtained from either scale alone. For this reason, the Full Scale IQ was used for comparative purposes with the individual Stanford-Binet IQ. The resulting correlation of .27 between the Stanford-Binet and the Full Scale is comparable to that of .28 reported by Rohrs and Haworth (1962) between the original DAM, which was comprised of a single Draw-a-Man score, and the 1960 Stanford-Binet using mentally retarded adolescents. Kennedy and Lindner (1964) found initial correlations from .29 to .41 with the same two instruments on Southeastern Negro school children. The correlation increased to .67 after the authors weighted the items on the DAM.

Correlations between the original DAM and earlier versions of the Stanford-Binet have yielded coefficients from .41 to .72 with mentally retarded adolescents and adults (Birch, 1949; Earle, 1933; Israelite, 1936; McElwee, 1932; Williams, 1929). These figures are somewhat higher than the present finding.

The modest degree of relationship shown between these two instruments in this study may confirm the hypothesis that the instruments measure somewhat different abilities. On the other hand, it should be pointed out that the Stanford-Binet IQs were determined from one to five years prior to this study; this time lag between testing and/or the restricted range of IQs could account for the low correlation reported.

Rohrs and Haworth (1962), studying mental retardates, found correlations between the DAM and WISC Performance Scale to be .53. Tobias and Gorelick (1960) reported a correlation of .50 between DAM and Wechsler Adult Intelligence Scale (WAIS) performance scores. This study also reported a coefficient of .63 between the original DAM and a worker efficiency rating scale although the WAIS performance scores seemed to be a better predictor of worker efficiency. Tobias and Gorelick conclude that since a higher correlation was obtained between the DAM and the WAIS Performance Scale than the DAM and the WAIS Verbal Scale, it appears that factors similar to those involved in other performance tests are required for achievement on the DAM. Using the original DAM and the WAIS, Gunzburg (1955) reported a correlation of .73 between the performance scale and the DAM scores for cultural familial mentally retardated adults. For the mentally retarded adolescent and pre-adolescent, the modest relationship between the Stanford-Binet and drawing IQs may be a reflection of the retardate's less than adequate verbal abilities which the Stanford-Binet emphasizes; hence, the drawing scores may indicate a somewhat greater correspondence to non-verbal tasks.

That there are differences in the drawing performance of mentally retarded children and those of normal children is well-established. The type of items on the drawing scale for which retarded Ss usually gain credit is generally more concrete or detailed while the more abstract components of the drawing task, e.g., spatial orientation, proportion of body parts, sketching technique, and depiction of motion, are those items for which relatively few retarded Ss gain credit (Earle, 1933; Goodenough, 1926; Israelite, 1936).

Consistent with the present findings, previous studies with educable mental retardates (mainly cultural-familials) have shown that scores on the DAM tend to run higher with this group than scores on the Stanford-Binet. Mitchell (1959) found that her Ss scored in the mildly retarded range (Binet IQs 52-67) on the original DAM although they had been evaluated as being moderately retarded (Binet IQs 36-51) by individual assessment. Other studies have reported higher DAM scores than individually administered Stanford-Binet scores for cultural-familial mental retardates (Birch, 1949; Rohrs & Haworth, 1962).

Mitchell (1959) suggested in her study that the maximum CA be extended upward in order to equalize DAM IQs with Stanford-Binet IQs, while Tobias and Gorelick

(1960) found that closer agreements between DAM and WAIS scores resulted from using CA 12 rather than CA 16 for scoring the DAM, implying that the DAM scores, when calculated by ordinary methods, were lower than WAIS scores. The revised version of the DAM has extended the maximum CA from 13 to 15, but the Ss in the present study, like those in Mitchell's, scored clearly above their Stanford-Binet ranges of mild to borderline (IQs 56-72). It is not clear which procedure should be used in arriving at drawing test IQs for older mentally retarded individuals. However, a more consistent and accurate method should be established if the drawing test is to be of continued use with this population. Their DAM IQs were in the borderline to average range (IQs 69-89) with a mean difference of 13.7 points, nearly one standard deviation. This discrepancy may be related to the abilities measured by the two instruments and may be applicable only to those Ss who are mildly retarded. Most of these Ss exhibit cultural-familial mental retardation. The discrepancy shown here has not been typical of all groups of mentally retarded persons. Gunzburg (1955) reported higher DAM scores for cultural-familial mentally retarded Ss than for organically damaged mental retardates. Thus, mentally retarded Ss may either be penalized by the verbal nature of the Stanford-Binet or aided by the measurement of concrete cognitive concepts on the drawing test.

Since there is no empirical evidence to indicate exactly what abilities are being assessed by the drawing test, Harris' statement (1963) that the drawing test is not more allied with performance than verbal items may not be accurate. In fact, the available evidence would suggest that for the mentally retarded, factors involved in the drawing test seem to correspond more closely with performance tasks.

Test-Retest Reliability

Although the seven-month interval and the restricted range of IQs (initial Full Scale IQs 74 to 86) would tend to decrease the size of the correlation, the coefficient of reliability between the Full Scale initial test and retest for the total sample (N = 343) was .81. This compares favorably with previously reported correlations of .77 to .91 for mentally retarded Ss (Brill, 1935; Yepsen, 1929).

Male Ss (N = 213) had a significantly lower test-retest reliability coefficient for the Man Scale (.69) than for the Woman Scale (.77), while no such differences existed for female Ss. No explanation for this significant difference for the male Ss is readily suggested. It may be noted that the male Ss increased in scores on the Man Scale from 80.1 on the initial test to 83.7 on the retest which resulted in a significant difference ($t = 2.21$, $p. < .05$) between test and retest scores.

The difference in reliability coefficients for the male Ss of .69 for the Man Scale and .80 for the Full Scale was significant at the .01 level. This finding for male Ss on the Man Scale is contrary to the statement by Harris (1963) that the statistical reliability of either scale alone is higher than that of the Full Scale. Moreover, the male Ss' reliability on the Woman Scale and the female Ss' reliability on both scales were not significantly different from the reliability of the Full Scale, further refuting the statement by Harris.

Alternate-Form Reliability

The relationship between the Man Scale and the Woman Scale on the initial test was .75 for the total sample (N = 343), and .81 on the retest, a difference which is significant at the .05 level. Although both the male and female Ss had higher alternate-form reliability coefficients on the final test than on the initial test, these differences were not statistically significant.

Harris (1963) reported a .75 alternate-form reliability coefficient for his standardization population, which is lower than his previously reported test-retest reliability correlation. He suggested that the Man and Woman Scales may evaluate somewhat different abilities; since it is not known what specific abilities are being measured by each scale, their use as parallel or substitute forms of the same test is not suggested.

Differences between Initial Test and Retest Standard Scores

The total sample (N = 343) had an average increase of 2.2 points on Full Scale standard scores after a seven-month interval. This increase was statistically significant ($t = 2.18$, $p < .05$).

Because Harris (1963) found sex differences in the performances of boys and girls on the two tasks of the scale, he established separate norms for male and female Ss on both the Man and Woman Scales. For this reason, it will be necessary to discuss score changes for each sex on the separate scales.

The male Ss (N = 213) had an average increase of 3.6 points on the retest scores of the Man Scale. Only CA 18 showed a decrease in points on the retest. CA groups 13, 15, 17, and 20 showed larger increases in scores than did any of the other CA groups.

The CA 17 male Ss had an increase of 5.2 standard score points on the Man Scale but a decrease of 2.2 points on the Woman Scale. The loss of points on the Woman Scale at this CA, as well as for the CAs 18 and 19, seems to indicate that for the total group of mentally retarded male Ss, the ceiling of the Woman Scale may be at CA 17; thus the test ceases to be an effective measure of further growth in mental maturity. On the other hand, the low Stanford-Binet IQ level of the CA 17 male Ss (56.3) plus the rather large increase on the Man Scale upon retest suggests that they are still elaborating concrete concepts of a man.

Because of the small Ns in the older CA groups (17, 18, 19), however, this finding may not be conclusive; differences did not reach statistically significant levels. Despite the small number, the five CA 20 male Ss showed a statistically significant increase ($p<.05$) of 10.6 points on the Man Scale and an increase of 12.6 points ($p<.05$) on the Woman Scale. This discrepancy with the trend of scores at the other age levels cannot be readily explained.

The female Ss (N = 130) showed a slight increase in mean retest scores on the Man Scale (1.2 points) and a negligible increase in mean scores on the Woman Scale (0.7 points), differences which did not reach statistical significance at the .05 level. There was a noticeable difference between the ages of the female Ss compared to the male Ss in terms of test ceiling. On the Man Scale, older female Ss as a whole showed a decrease in points at CA 16 (N = 8) and at CA 18 (N = 5). The CA 18 group decreased 7.0 standard score points, the largest change to occur for female Ss in either direction. No female group above age 15 gained more than a point between tests, even though for this group, because of the CA 15 divisor, no adjustment for age was made.

The female Ss' performance on the Man Scale showed very little change between tests, but the decrease in standard scores on the retest of 1.2 points at CA 16 and the 0.7 point loss at CA 18 again possibly indicates the faster maturation and development of girls and perhaps the earlier age (compared with males) at which the Man Scale ceases to be a useful assessor of continued growth in mental maturity. The comparability of score changes on both scales suggests that the sex role identification of female Ss may be stronger than that of their male peers, or as Lowenfeld and Brittain (1964) suggest, the interest in the opposite sex is seen at an earlier age in female's drawings than in those of males. For this reason, either scale alone may be an adequate assessor of the abilities tapped by the test for EMR female Ss up to CA 16.

Conclusions

1. The intra-scorer reliability of .99 on the re-scoring of 25 drawings of a man and 25 drawings of a woman indicated that the test can be reliably scored.

2. Although the results seem to indicate that the test is reliable for use with a retarded population, it is not safe to say that the test would be as reliable with other populations.

3. It was also concluded that the DAM test is useful with educable mentally retarded male Ss through CA 20. The test may be somewhat less useful with older female Ss, who ceased to increase scores, after a seven-month interval, at CAs 16 and above.

4. The maximum CA divisor established by Harris (1963)

appears to be appropriate for mentally retarded Ss, and the drawing test as a measure of concrete concept formation seems to be a reliable instrument for gaining information about mildly retarded adolescents, at least to supplement other types of psychological data. The results of the study discussed here would tend to confirm Birch's conclusion (1949) that for older mentally retarded adolescents (IQ below 70), the drawing test appears to be a useful instrument, although questions concerning its validity for specific purposes remain unanswered.

Implications

The following leads for future research are suggested by the present study:

(1) Since the two scales seem, in part, to measure different abilities, a factor analysis of the items of both the Man Scale and the Woman Scale together could be made to determine the inter-relation of items on the two scales.

(2) An analysis of the order of difficulty of items should be computed in order to compare the responses of educable mentally retarded Ss with those Ss used in standardization of the DAM.

(3) Analysis of the data in terms of racial differences, number and sex of adults in the family, and socio-economic variables would yield interesting data concerning the role of experience in determining scores on the scale.

(4) An attempt might be made to utilize the concept of mental age rather than the deviation IQ in analyzing the results of the drawing test for retarded individuals.

(5) Since the coefficient of correspondence between the 1960 Stanford-Binet and the Harris-Goodenough Drawing Test was only .27, a comparative study using another instrument emphasizing performance rather than verbal ability may reveal a closer agreement between DAM and performance scale IQs. It is suggested that in such a study individually administered intelligence test scores be obtained at the time of the DAM scores to maintain standard testing conditions.

CHAPTER V

SUMMARY OF THE STUDY

The Problem

The reliability of the performance of retarded children and adolescents on the revised version of the Draw-a-Man test (Harris, 1963) had not been explored, and it was not clear whether the decline in performance which occurs in normal adolescents in mid-teens also typifies the performance of retarded individuals. Mitchell (1959) gave some evidence that increases in test scores continued throughout adolescence on the original test formulated by Goodenough (1926) and suggested that evidence from tests of intelligence other than the drawing test clearly indicated that mental age in mentally retarded persons continues to increase well beyond thirteen years. The age range on the Harris-Goodenough Drawing Test has been extended to fifteen years. If the drawing test yields an adequate MA, this MA would be expected to increase at about the same rate and for the same duration of time as MAs derived from other intelligence examinations, provided that the test ceiling had not been reached. If the drawing test scores do increase after CA 15-0, IQs operations level.

Review of the Literature

The Goodenough test scores of normal children cease to show increments soon after Bayley's manipulating symbols period of mental development terminates and during Piaget's shift from the period of concrete operations to the period of formal operations. This suggests that the drawing test evaluates primarily the ability to form concrete concepts (Harris, 1963). Undoubtedly, the concept of a person as a concrete object undergoes an elaborate differentiation with age. As the child moves into Piaget's period of formal operations, his intellectual processes are sufficiently advanced and complex to allow him to conceptualize abstract and hypothetical relationships as well as concrete ones. Governed by the rules of logic, his thinking now characteristically involves higher order abstractions. Since it taps more concrete concepts, the drawing test at this time ceases to show increments and therefore ceases to be an index to the child's further growth in intellectual maturity (Harris, 1963).

To date Mitchell (1959) has conducted the largest study with retarded Ss. She used 536 institutionalized Ss, and found that the raw scores and the MAs which they represent continued to increase fairly rapidly through CA 15. She also found that half of her sample of moderately retarded Ss scored in the mildly retarded range in the 15 and 16 year old groups on the drawing test, a situation clearly out of line with their intellectual capacities as measured by individual inteligence tests.

Rohrs and Haworth (1962) found that mean scores on the Stanford-Binet and the Draw-a-Man tests were more nearly comparable for retarded Ss than for normals, who tended to score higher on the Stanford-Binet. In contrast, Kennedy and Lindner (1964) found that Negro children in the Southeastern United States scored somewhat higher on the drawing test than on the Stanford-Binet, probably because of the highly verbal nature of the Stanford-Binet.

Objectives

This study was undertaken to determine the most appropriate CA divisor to use with children whose mental age rendered them suitable candidates for the drawing test, but whose chronological age was greater than that of the Ss in Harris' restandardization population.

The following questions were asked in the experiment:

(1) Can the test be reliably scored?

(2) Are the IQs obtained on the Harris-Goodenough Drawing Test stable over a seven-month interval? Is there variability by age and sex in such retest stability?

(3) At what age, if any in this range, do the test scores obtained by educable mentally retarded adolescents

cease to increase on retest after a seven-month interval?

(4) What is the most appropriate CA to employ in calculating the IQs of such Ss above the presently recommended maximum age of 15-0?

Procedures

Sample.--A total of 572 Ss from two Piedmont North Carolina communities was randomly selected from the entire enrollment of intermediate and secondary educable mentally retarded classes and was grouped according to CA on initial test administration. Ss ranged in CAs from 11-0 to 20-6 and had Stanford-Binet IQs from 56 to 72. At the time of the retest, only 343 Ss of the original sample were still enrolled in special classes. These 343 Ss constituted the final sample.

Methods used.--The Harris-Goodenough Drawing Test was administered in group form to all EMR classes in Durham and Greensboro, North Carolina, at the intermediate and secondary levels. The test was readministered after a seven-month interval. It was administered in the Ss' own classrooms by the classroom teacher under written directions of the investigator. A single trained examiner, unaware of the age or previous Stanford-Binet IQ of any S, which was secured from confidential school records, scored all the drawings. Therefore, any error in scoring was presumably consistent.

of mentally retarded adolescents might be overestimated by an artificial restriction of their chronological age increment. A gradual and spurious increase in IQ from CA 15-0 to that point at which MA growth ceased would be noted.

During adolescence, progress in drawings made by normal individuals becomes laborious and slow, and often shows deterioration or regression ascribed to emotional conflict present after puberty. There is an increased power of observation, and although cognitive and intellectual functions are present, there is usually also a critical self-awareness. The drawing test is thus not very useful with children older than twelve or thirteen with normal or above average intelligence (Harris, 1963). Harris also suggested that adolescents become involved in sketching. This new attempt at abstraction, which is perhaps related to Piaget's period of formal operations, results in lowered scores on the test which apparently measures primarily the ability to form concepts of the type developed during the Piagetian concrete operations stage of development.

Mitchell (1959), however, found that the original Draw-a-Man test continued to prove useful with older retarded adolescents or adults who may or may not manifest this critical self-awareness attitude, or in whom the appearance may be delayed, or those retardates who have not made the shift from the concrete to the formal

Results

Statistical procedures used in analyzing the data provided these answers to the questions proposed by the investigator:

(1) The intra-scorer reliability coefficient of .99 after a six-week interval on both the Man Scale and the Woman Scale indicated that the test can be scored reliably.

(2) The over-all mean change for the 343 Ss was an increase of 2.2 points ($p < .05$) between test and retest. Test-retest reliability was .81 ($p < .01$). Analysis of variance yielded significant F-ratios which indicated that the standard deviations of the change differed at the various CA groups for male Ss.

(3) On the Harris-Goodenough Drawing Test, scores continued to increase through CA 16 for the total sample of EMR Ss. At CA 17, there was no statistically significant difference between test and retest IQs on the Full Scale. The decrease in scores on the retest was evident at CA 18. However, when analyzed by sex, male Ss in all CA groups except CA 18 continued to increase on the Full Scale drawing test. Female Ss' scores decreased at CA 16 and above.

(4) Harris' present maximum CA divisor of 15 seems to be adequate for mentally retarded adolescents although the drawing test continues to be useful for male Ss through

CA 20; whereas, for female Ss the test becomes a less accurate index of the abilities measured at CA 16 and above.

Implications for Future Research

The following conclusions and implications were drawn from the study:

(1) A system of score interpretation using MAs as well as deviation IQs might furnish meaningful information about the performance of EMR children.

(2) An item analysis of the present data would yield those items passed by EMR (largely cultural-familial) Ss compared with items passed by Ss of normal intelligence, equivalent in over-all raw score.

(3) A factor analysis should be computed on the Man Scale and the Woman Scale together to determine the inter-relationship of the two scales.

(4) Since the validity coefficient of .27 between the 1960 Stanford-Binet and the Harris-Goodenough Drawing Test may be a reflection of the highly verbal nature of the Stanford-Binet, a comparative study using an instrument which incorporates some of the performance aspects measured by the drawing test may reveal a closer agreement. It is suggested that individually administered intelligence test scores be obtained at the time of the DAM scores to maintain standard testing conditions.

APPENDIX A

GREENSBORO PUBLIC SCHOOLS
Greensboro, N. C.

TO:

FROM: Irwin S. Levy
Principal Investigator

DATE: April 15, 1966

Enclosed you will find envelopes of materials to be directed to the Special Education Teachers whose names appear on the outside. Within each envelope there is a letter of instructions, and general statements for clarification of purpose and procedures for the use of the materials.

The Greensboro City Administration Unit has been chosen to be a part of an experimental study which has as its purpose to determine the reliability of the performance of educable mentally retarded children on the Harris-Goodenough Drawing Test. The effect of age as a variable in the standardization of this test has not been explored. The failure to take this into account has produced a questionable effect of the reliability of this test, and our cooperation with this effort can make a major contribution to education in general.

Mr. Weaver is in complete agreement to our being a part of the study, so we will appreciate your cooperation in passing these envelopes to the teachers whose names appear on the outside. If there are any questions, please feel free to telephone.

The amount of time which will be taken up in handling the materials, provided herein, would at the most be no more than 10 or 15 minutes per teacher. The materials should be back in our office by April 27th in order for us to have it in the hands of the investigators on April 29th.

DURHAM CITY SCHOOLS
Durham, N. C.

TO:

FROM: Irwin S. Levy
Principal Investigator

DATE: April 5, 1967

Enclosed you will find envelopes of materials to be directed to the Special Education Teachers whose names appear on the outside. Within each envelope there is a letter of instructions and general statements for clarification of purpose and procedures for use of the materials.

The Durham City Schools has been chosen to participate in an experimental study which is being supported by the U.S. Office of Education. The test is the same which you administered to the students this fall and will add additional information of assessment to their permanent folders. Your cooperation with this effort can make a major contribution to education in general.

Miss Lipscomb is in complete agreement to your being part of the study, so we will appreciate your cooperation in passing these envelopes to the teachers whose names appear on the outside. The amount of time which will be taken up in handling the materials provided herein would, at the most, be no more than ten or fifteen minutes.

In order to test all students whose test sheets are in the envelope, it will not be necessary to return the envelope to Miss Lipscomb's office until Friday, April 21, 1967. Thank you for your cooperation.

APPENDIX B

INSTRUCTIONS

Enclosed you will find two sheets of blank paper bearing the names of the children who are to be included in the study.

> NOTE: It is not necessary that all children be tested at the same time nor on the same day nor is the test to be timed. Usually five to eight minutes is enough for the two exercises. Have each child use a No. 2 pencil.

EXERCISE I - TO BE DONE ON ONE PAGE GIVING THE CHILD'S NAME -

> TEST ADMINISTRATOR SAYS: Draw a picture of a MAN, the very best MAN you can draw. Be sure you draw the WHOLE MAN, not just the head.

EXERCISE II - TO BE DONE ON THE SECOND PAGE GIVING THE CHILD'S NAME -

> TEST ADMINISTRATOR SAYS: Draw a picture of a WOMAN, the very best WOMAN you can draw. Be sure you draw the WHOLE WOMAN, not just the head.

When both tests have been completed, make no marks on the sheets or supply any written comments. Place the sheets back in the envelopes and return the envelope to your Principal in order that he may forward them to the Director of Special Education.

REFERENCES

Bayley, Nancy. A new look at the curve of intelligence. In Proceedings, 1956 Invitational Conference on Testing Problems, Educational Testing Service, Princeton: 1957. Pp. 11-23.

Birch, J. W. The Goodenough drawing test and older mentally retarded children. Amer. J. ment. Defic., 1949, 54, 218-224.

Brill, M. The reliability of the Goodenough Draw-a-Man Test and the validity and reliability of an abbreviated scoring method. J. educ. Psychol., 1935, 26, 701-708.

Buhler, K. The mental development of the child. London: Routledge and Kegan Paul, Ltd., 1930.

Earl, E. J. C. The human figure drawings of adult defectives. J. ment. Sci., 1933, 79, 305-328.

Edwards, A. L. Statistical methods for the behavioral sciences. New York: Holt, Rinehart and Winston, 1964.

Ferguson, G. A. Statistical analysis in psychology and education (2nd ed.). New York: McGraw-Hill, 1966.

Goodenough, Florence L. Measurement of intelligence by drawings. New York: Harcourt, Brace & World, 1926.

Gunzburg, H. C. Scope and limitations of the Goodenough drawing test method in clinical work with mental defectives. J. clin Psychol., 1955, 11, 8-15.

Harris, D. B. Children's drawings as measures of intellectual maturity: a revision and extension of the Goodenough Draw-a-Man Test. New Ylrk: Harcourt, Brace & World, 1963.

Israelite, Judith. A comparison of the difficulty of items for intellectually normal children and mental defectives on the Goodenough drawing test. Amer. J. Orthopsychiat., 1936, 6, 494-503.

Kennedy, W. A. & Lindner, R. S. A normative study of the Goodenough Draw-a-Man Test on southeatern Negro elementary children. Child Develpm., 1964, 35, 33-62.

Lowenfeld, V., & Brittain, W. L. Creative and mental growth (4th ed.). New York: Macmillan, 1964.

McCarthy, Dorothea. A study of the reliability of the Goodenough drawing test of intelligence. J. Psychol., 1944, 18, 201-216.

McElwee, Edna W. The reliability of the Goodenough intelligence test used with sub-normal children fourteen years of age. J. appl. Psychol., 1932, 16, 217-218.

Mitchell, Anna Carr. A new maximum CA for the Draw-a-Man Test. J. consult. Psychol., 1959, 23, 555-557.

Piaget, J. The origin of intelligence in the child. London: Routledge and Kegan Paul, 1953.

Robinson, H. B. & Robinson, Nancy M. The mentally retarded child: a psychological approach. New York: McGraw-Hill, 1965.

Rohrs, F. W. & Haworth, Mary R. The 1960 Stanford-Binet, WISC, and Goodenough tests with mentally retarded children. Amer. J. ment. Defic., 1962, 66, 853-859.

Sundberg, N. D. The practice of psychological testing in clinical services in the United States. B.P.S. Bull., 1961, 44, 1-9.

Terman, L. M. & Merrill, Maud A. Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin, 1960.

Tobias, J. C. & Gorelick, J. The utility of the Goodenough scale in the appraisal of retarded adults. Amer. J. ment. Defic., 1960, 65, 64-68.

Vinacke, W. E. Concept formation in children of school ages. Educ., 1954, 74, 527-534.

Williams, J. H. Validity and reliability of the Goodenough intelligence test. Sch. & Soc., 1935, 41, 653-656.

Yepsen, L. N. The reliability of the Goodenough drawing test with feebleminded subjects. J. educ. Psychol., 1929, 20, 448-451.