Spring 1972

REALIABILITY OF KOPPITZ'S EMOTIONAL INDICATORS ON THE HUMAN FIGURE DRAWINGS OF CHILDREN

Marcus J. Miles

REALIABILITY OF KOPPITZ'S EMOTIONAL INDICATORS ON THE HUMAN FIGURE DRAWINGS OF CHILDREN

A Thesis

Presented to

the Graduate Faculty

Central Washington State College

In Partial Fulfillment

of the Requirements for the Degree

Master of Education

---

by

Marcus J. Miles

May 1972

ACKNOWLEDGMENTS

The author wishes to thank the following people for their assistance with this project: George Beckstead, Colin Condit, Marlene Jensen, Jim Klahn, Merle Palmerton, Travis Rundell, Ron Zirker, and Max Zwanziger. Several other people, not listed here, played a part in moving this work toward completion. Their efforts were appreciated. Finally, a special thanks must go to Marjie, for the countless hours of hard work she contributed.

TABLE OF CONTENTS

LIST OF TABLES

RELIABILITY OF KOPPITZ'S EMOTIONAL INDICATORS ON THE HUMAN FIGURE DRAWINGS OF CHILDREN

by

Marcus J. Miles

May, 1972

The study was designed to investigate inter-scorer consistency and test-retest reliability of Koppitz's "Emotional Indicator" scoring system for human figure drawings.

Drawings were secured from 438 third grade pupils. A test-retest interval of one month was employed.

Phi coefficients ranging from .75 to .81 were found for inter-scorer consistency between three judges. Twenty-eight of Koppitz's 30 "Emotional Indicators" and her all important total score category yielded test-retest phi coefficients smaller than .43.

CHAPTER I

INTRODUCTION

The purpose of this study was to investigate the reliability of Koppitz's Emotional Indicators on the human figure drawings (HFDs) of third grade children.

In 1968 Elizabeth Koppitz presented the results of several investigations designed to develop and validate an objective method for scoring and analyzing the HFDs of children age 5 to 12 years. One of the major results of her work was a list of 30 signs which she termed Emotional Indicators.

1. Poor integration	11. Crossed eyes	21. Three figures
2. Shading face	12. Teeth	22. Clouds
3. Shading body	13. Short arms	23. No eyes
4. Shading hands	14. Long arms	24. No nose
5. Asymmetry	15. Clinging arms	25. No mouth
6. Slanting figure	16. Big hands	26. No body
7. Tiny figure	17. Hands cut off	27. No arms
8. Big figure	18. Legs together	28. No legs
9. Transparency	19. Genitals	29. No feet
10. Tiny head	20. Monster	30. No neck

To qualify as an Emotional Indicator each sign had to meet three criteria.

1. It must have clinical validity, i.e., it must be able to differentiate between HFDs of children with and without emotional problems.
2. It must be unusual and occur infrequently on the HFDs of normal children who are not psychiatric patients, i.e., the sign must be present on less than 16% of the HFDs of children at a given age level.
3. It must not be related to age and maturation, i.e., its frequency of occurrence on HFDs must not increase solely on the basis of the children's increase in age (Koppitz, 1968, p.4).

The clinical validity, criteria #1, of each sign was determined by comparing the drawings of 76 matched pairs of clinic patients and "all around" students. Indicators which occurred more frequently or with equal frequency on the drawings of "all around" students were discarded. Signs which did not meet criteria #2 (frequency of occurrence less than 16% on the drawings of normal children) and criteria #3 (frequency of occurrence unrelated to age and maturation) were eliminated in a normative study employing the drawings of 1856 elementary school children.

It was not the purpose of the present investigation to assess the validity of Koppitz's three criteria or to scrutinize the methods by which she applied them. Rather, it was the purpose of this study to suggest and investigate a fourth criteria, reliability. According to the American Psychological Association,

> The test manual should indicate to what extent test scores are stable, that is, how nearly constant the scores are likely to be if a test is repeated after time has lapsed (French and Michael, 1966, p. 30).

Koppitz, in her recent volume, neglected to report test-retest reliability data. The primary goal of the present investigation was to supply this missing information.

A secondary purpose was to gather and evaluate inter-scorer consistency data. Evidence of this type of reliability was included in the Koppitz manual but was presented in terms of percentage of agreement, a practice questioned by Swensen.

> Serious criticism must be leveled against the use of the percentage of agreement as a measure of reliability. The significance of the percentage of agreement on the DAP is entirely dependent upon the base rate of the particular body part or structural aspect of the drawing that is being investigated (1957, p. 434).

Koppitz did not consider base rates when evaluating her data. This fact, coupled with Swensen's criticisms, indicated a need for further study of inter-scorer consistency.

Prior to the present investigation several studies of stability and inter-scorer reliability had been made of various scoring methods and drawing signs similar to the signs and methods advocated by Koppitz. Studies concerned with test-retest reliability will be discussed first.

Nichols and Strumpfer (1962) obtained male and female HFDs from 197 adult subjects (Ss). Drawings were scored for the presence or absence of 14 drawing details. The stability of these details was estimated by scoring the male and female drawings separately for each detail and calculating phi coefficients as an index of reliability over

the two drawings. Of the 14 signs studied, nine bore no resemblance to any of Koppitz's Emotional Indicators. The five Koppitz-like signs investigated, transparency, lack of body part, shading, figure off balance, and figure very small, earned phi coefficients of .26, .51, .31, .38, and .51 respectively.

Incompletions and height of figure were among seven drawing signs investigated by Starr and Marcuse (1959) in a study employing 193 male and female college students as Ss. Test-retest intervals of one month and immediate were used. The stability of incompletions over two drawings was measured by computing the correlation statistic, phi/phi maximal. The result was a coefficient of .54. A product-moment coefficient of .23 was reported for the sign, height of figure.

A product-moment r of .61 was reported by Bradshaw (1952), in a study cited by Swensen (1957), for the sign, height of figure. Bradshaw, using a one week interval, administered the HFD Test to 100 male and female college students.

Height of figure was also investigated by Apfeldorf, Randolph and Whitman (1966). A total of 51 institutionalized veterans were asked to draw a human figure. Upon completion of the first drawing a second was immediately requested. Comparison of first and second drawings resulted in a product-moment r of .88.

Hammer and Kaplan (1964c, 1966) and Lehner and Gunderson (1952) studied a host of signs similar to Koppitz's

scoring categories, but neglected to report the degree to which signs were found reliable, that is, the correlation between test-retest scores. As a consequence, data from these studies resists meaningful interpretation and has been omitted from this discussion. Studies (Guinan and Hurley, 1965; Hammer and Kaplan, 1964a, 1964b; Litt and Margoshes, 1966; Strumpfer, 1963) concerned with the test-retest reliability of HFD signs other than those employed by Koppitz have also been omitted from this review.

No investigations prior to the present study, had attempted to measure inter-scorer consistency for Koppitz's Emotional Indicators exclusive of other scoring systems. Koppitz's normative study, criticized earlier for presenting data in terms of the percentage of agreement between scorers, was additionally flawed when scoring of Emotional Indicators was combined with scoring of Developmental Items before inter-rater reliability was determined. Likewise, Hall and Ladriere (1970) rendered the results of their study useless for purposes of the present study by combining scoring of Koppitz's Emotional Indicators with scoring by five other systems before calculating inter-scorer agreement.

As in the case of test-retest stability studies, several scorer-consistency studies (Craddick, Leipold and Cacavas, 1962; Grams and Rinder, 1958; Handler and Reyher, 1964; Hoyt and Baron, 1959; Mogar, 1962; Solar, Bruehl and Kovacs, 1970; Sopchak, 1970; Strumpfer, 1963) have been

omitted from this discussion since they investigated scoring methods dissimilar to Koppitz's system.

As is no doubt obvious from the foregoing discussion, test-retest stability and inter-scorer consistency for Koppitz's scoring scheme could not be predicted from the results of previous research. No usable data concerned with inter-judge reliability was available and findings reported for the individual stability of signs similar to Koppitz's Emotional Indicators were fragmentary and incomplete.

Thus came the impetus for the present investigation.

CHAPTER II

METHOD

Subjects

A total of 438 third grade pupils, 240 boys and 198 girls, served as subjects (Ss) for this investigation.

Originally, 477 third grade pupils (all children enrolled in regular third grade classrooms in Moses Lake, Washington, School District during the months of April and May, 1970) were designated as Ss, but 39 were unable to complete two scorable drawings, either because of absence during testing or because of failure to complete a finished drawing within the 15 minute time limit.

Materials

Materials used in this study included: 200 "Pueblo" brand #2 pencils, 1000 sheets of 8½ by 11 inch white paper, 80 standard manila folders stapled together in pairs to form desk dividers, 21 large manila envelopes, 21 letters of explanation for classroom teachers (Appendix A), two copies of Koppitz's test manual, Psychological Evaluation of Children's Human Figure Drawings, a stop watch, a ruler, and a protractor.

Procedure

Prior to the collection of data, a teacher's aide was instructed in test procedures by the experimenter (E). The teacher's aide then practiced those procedures by administering the HFD Test to a class of second graders and a class of fourth graders in E's presence. Following the practice sessions E and the teacher's aide discussed problems encountered and devised a set of standard test procedures to be followed (Appendix B). Within the set of standard procedures Koppitz's specific instructions for group administration of the HFD Test were adhered to with one exception, addition of a 15 minute time limit.

During the third week in April and again during the third week in May the teacher's aide administered the HFD Test in 21 third grade classrooms. E was not present during these sessions. Testing was completed in five days and was conducted between the hours of 12:30 p.m. and 3:15 p.m. The time limit for each session did not include time used for giving of instructions and collection of drawings. Desk dividers were placed between students seated at double desks to minimize the possibility of copying.

A large manila envelope was prepared in advance for each classroom. On the back of each envelope was listed the name of the school, the teacher's name, and the names and birthdates of all pupils enrolled in that classroom. Adjacent to each student's name was placed a black code number

and a red code number. Code numbers were selected from a table of five digit random numbers (Snedecor, 1967). In April, as drawings were collected the black code number listed for each child was placed on his or her drawing. In May, the appropriate red code number was placed on each drawing.

Once all drawings had been collected and coded, names were removed and thorough shuffling was conducted by the teacher's aide. Drawings were then presented to E for scoring.

Scoring was conducted by E according to Koppitz's instructions. All 976 drawings were scored for the presence or absence of Koppitz's 30 Emotional Indicators. The scoring process involved looking at each drawing and recording for that drawing which signs were present. For example, one of the drawings scored by E received a score of 2, 15, 22, indicating the presence of three signs, shading of face, arms clinging to body, and clouds. Scoring of all drawings was completed in approximately 90 hours over a period of several weeks.

A certified school psychologist, unfamiliar with Koppitz's scoring system, was then asked to score 50 of the same drawings selected at random. Scoring was conducted in two sessions. During the first session the psychologist was given a copy of Koppitz's test manual, a ruler, a protractor, and 25 drawings, and was asked to score the drawings as best

he could without further discussion. Prior to the second session, held immediately following the first, E and the psychologist discussed Koppitz's scoring system for approximately 20 minutes. An attempt was made to agree upon scoring interpretations. The psychologist was then asked to score the remaining 25 drawings.

A fifth grade teacher, also unfamiliar with Koppitz's scoring scheme, scored the same 50 drawings under similar conditions.

All drawings were then decoded. April drawings were paired with their May counterparts and scores were recorded on a master list. Data was separated into three categories in conformity to the three types of findings sought, test-retest reliability of individual Emotional Indicators, stability of total score, and inter-rater reliability. Each set of data was found to be dichotomous in nature. Emotional Indicators had been scored as present or absent, total score for individual drawings had been scored as two or more signs present versus one or no sign present, and consistency of scoring between judges had been tabulated as agreement versus disagreement.

A statistic for dealing with dichotomous data, both genuine and artificial, was suggested by Garrett.

> When the classification is truly discrete and the variables can take only one of two values, the phi coefficient is an appropriate measure of correlation. Phi may be used also with continuous variables which have been grouped into two categories, . . . (1958, p. 389).

Following Garrett's suggestion, the reliability of each Emotional Indicator, the stability of total score, and inter-judge consistency were determined by calculating phi coefficients. The significance of all phi coefficients was determined by computing chi squares.

The statistic phi/phi maximal used by Starr and Marcuse (1959) was not employed. According to Guilford,

> It is recommended that the maximal phi that suits any given situation be considered when interpreting an obtained phi as representing a strength of the intrinsic relationship between two variables. The word intrinsic is stressed here, because the actual size of phi indicates the degree of practical, predictive value of the relationship (1956, p. 314).

The goal of the present study was to produce data of predictive value.

CHAPTER III

RESULTS

The purpose of this study was to gather and evaluate test-retest and inter-rater reliability data for a new HFD scoring system devised by Elizabeth Koppitz.

Inter-scorer reliability data was obtained by asking a school psychologist and a fifth grade teacher, both initially unfamiliar with Koppitz, to score 50 drawings previously scored by E. Each scored 25 drawings without practice or instruction from E. Results of this initial scoring by inexperienced judges are presented in Table 1. A second set of 25 drawings was scored by the same judges following 20 minutes of instruction from E. Results of this scoring are presented in Table 2.

TABLE 1

INTER-SCORER RELIABILITY FOR KOPPITZ'S EMOTIONAL INDICATORS PRIOR TO TRAINING OF JUDGES BY E

Scorer	Agree	Dis-agree	Phi	X^2	P
E vs. psychologist	696	54	.51	200.47	.01
E vs. teacher	703	47	.58	249.67	.01
Psychologist vs. teacher	678	72	.41	126.83	.01

TABLE 2

INTER-SCORER RELIABILITY FOR KOPPITZ'S EMOTIONAL INDICATORS FOLLOWING TRAINING OF JUDGES BY E

Scorer	Agree	Dis-agree	Phi	X^2	P
E vs. psychologist	730	20	.78	457.47	.01
E vs. teacher	733	17	.81	486.02	.01
Psychologist vs. teacher	727	23	.75	425.26	.01

The scoring of 25 drawings for the presence or absence of 30 signs requires 750 scoring decisions. Thus each phi coefficient in Tables 1 and 2 represents the degree to which two scorers agreed in making 750 such decisions.

Test-retest reliability was calculated for each of Koppitz's 30 Emotional Indicators and for total score. Koppitz divided the 30 signs she termed Emotional Indicators into three categories; Quality Signs, Special Features, and Omissions. Stability findings for the nine Quality Signs are presented in Table 3. Special Features are represented in Table 4. Table 5 presents findings for the eight Omission signs.

As a group, Quality Signs were found more reliable than Special Features. Omissions were found least reliable. The sign, no body, was found to be the most reliable of the Emotional Indicators investigated, with a phi coefficient of .71. The Indicator, monster or grotesque figure, was close

behind with a coefficient of .70. Phi coefficients ranging in size from .42 to -.02 were found for the remaining 28 Emotional Indicators investigated.

TABLE 3

TEST-RETEST RELIABILITY OF KOPPITZ'S QUALITY SIGNS

Emotional Indicator	N	Phi	X^2	P
Poor integration	438	.29	37.34	.01
Shading face	438	.13	07.65	.01
Shading body	384	.20	15.00	.01
Shading hands	437	.14	08.57	.01
Asymmetry	438	.06	01.68	.20
Slanting figure	438	.25	27.07	.01
Tiny figure	438	.22	21.59	.01
Big figure	435	.40	70.30	.01
Transparency	438	.29	37.86	.01

Total score for a single HFD is defined by Koppitz as the number of Emotional Indicators present. The presence of two or more Emotional Indicators on the HFD of a child is, according to Koppitz, suggestive of serious emotional problems. The stability of this all important diagnostic scoring category over two drawings for the 438 Ss employed in this investigation was found to be .21, as measured by phi. The emphasis Koppitz places on the use of this scoring category makes this the single most important finding of the present investigation.

TABLE 4

TEST-RETEST RELIABILITY OF KOPPITZ'S SPECIAL FEATURES

Emotional Indicator	N	Phi	X^2	P
Tiny head	438	-.01	00.01	.95
Crossed eyes	438	.42	79.08	.01
Teeth	438	.21	19.87	.01
Short arms	438	.26	30.18	.01
Long arms	438	-.02	00.18	.50
Clinging arms	438	.34	50.75	.01
Big hands	438	-.01	00.01	.95
Hands cut off	438	.15	09.56	.01
Legs together	438	.29	37.70	.01
Genitals	438	.12	05.93	.02
Monster	438	.70	217.51	.01
Three figures	438	.00	00.00	1.00
Clouds	438	.19	15.23	.01

TABLE 5

TEST-RETEST RELIABILITY OF KOPPITZ'S OMISSION SIGNS

Emotional Indicator	N	Phi	X^2	P
No eyes	438	-.01	00.02	.90
No noses	438	.34	49.83	.01
No mouth	438	-.01	00.07	.80
No body	438	.71	218.71	.01
No arms	438	-.01	00.01	.95
No legs	438	-.01	00.01	.95
No feet	386	.25	24.59	.01
No neck	185	.30	16.45	.01

Inspection of Tables 3 and 5 reveals that five Emotional Indicators have Ns smaller than 438, the number of paired drawings collected for this investigation. There is a simple explanation. The five signs involved are invalid, according to Koppitz, for use with certain age children. Hence, drawings by certain age Ss had to be removed from the sample before statistical analysis for these signs could be completed. An example may clarify the picture.

The Emotional Indicator, big figure, is defined by Koppitz as invalid for children seven years of age and younger. Three of the Ss employed in this study were seven years of age. Consequently, their drawings could not be used for estimating the stability of the sign, big figure.

As a result, N for this Indicator was reduced from 438 to 435.

The Emotional Indicators, shading of body, shading of hands, no feet, and no neck, were similarly affected when varying age restrictions set down by Koppitz (1968, p. 333-334) were applied.

CHAPTER IV
DISCUSSION

This chapter includes discussion of the following topics: limitations of the study, comparison of results with previous findings, problems encountered using Koppitz's scoring system, implications of the results, and suggestions for further research.

Limitations of the Study

Lack of experimental control was the major limitation of the present study. Because the study was conducted outside of the laboratory, results were subject to influence by extraneous variables. Testing was conducted in 21 classrooms, in seven schools, on different days, at different times during the afternoon. All testing was administered by the same examiner, but a different teacher was present in each classroom. The behavior of the examiner, of the teacher present, and of individual Ss undoubtedly varied from test session to test session, as did room size, temperature level, and so on. However, to E's knowledge, this lack of control did not produce any overwhelming problems. Attempts were made to control several important variables. Retesting of each group of Ss was done exactly one month

after testing, in the same room, by the same examiner, at the same time of day. A standard set of test procedures was followed, teachers and Ss were instructed as to how they should behave, and drawings were coded to avoid possible scorer bias.

A second limitation of the present study was the age and social background of the population studied. Third grade pupils from Moses Lake, Washington, a rural community, served as Ss. Most (90%) were eight and nine years of age, none was five, six, or twelve; and seven, ten, and eleven year olds comprised only 10% of the total population studied. Generalization of results is difficult from such a homogeneous population.

A third limitation was imposed by testing Ss in groups rather than individually. Koppitz recommended individual administration whenever possible. Group administration precluded the possibility of completely accurate scoring since Ss could not be questioned about unclear or ambiguous features of their drawings.

Comparison of Results with Previous Findings

The results of this investigation and the results of previous investigations, in the few cases where direct comparison was possible, were found to be generally consistent. During the present investigation phi coefficients of .29, .25, and .22 were found for the Emotional Indicators:

transparency, slanting figure, and tiny figure. Nichols and Strumpfer (1962) reported phi coefficients of .26, .38, and .51 for the same signs, but with adult Ss.

The stability of omissions ranged from .71 to -.02, as measured by phi in the present study. Starr and Marcuse (1959) found incompletions reliable to the extent of .54, as described by the statistic phi/phi maximal. Nichols and Strumpfer (1962) reported a phi coefficient of .31 for the sign, lack of body part.

Comparison of the inter-scorer reliability data gathered in the present study with previous results was not possible since no studies had before attempted to investigate inter-scorer consistency for Koppitz's Emotional Indicators, without first combining the system with other scoring methods.

Problems Encountered Using Koppitz's Scoring System

The present study was designed to study the reliability of a new scoring system for the HFD Test developed by Elizabeth Koppitz. Consequently, the Koppitz manual (1968) was used as a guide for scoring the HFDs collected. No report of the present study would be complete without mention of the problems encountered in using this manual.

While many of Koppitz's Emotional Indicators, such as tiny figure (figure two inches or less in height), are easily and accurately scored, a few are not. For example,

the sign short arms is defined as, "Short stubs for arms, arms not long enough to reach waist" (Koppitz, 1968, p. 332). The phrase "short stubs for arms" is not explained and no mention is made of how to score arms which fail to reach the waist because the waistline appears misplaced.

At least five of Koppitz's 30 Emotional Indicators are defined in such a way as to require subjective interpretations on the part of the scorer. Does shading hair count as shading? Do long arms caused by a short body count as long arms? Does a large shaded ring count as shading of hands? And so on. Scoring examples are provided in the form of actual drawings reproduced in miniature. Many, however, appear to show conflicting scoring approaches. Whenever scoring problems of this sort were encountered, each scorer decided how he would score that particular item on future drawings and then proceeded. Short arms due to a misplaced waistline, long arms caused by a short body, shading of hair, and large shaded rings were not scored.

The ambiguity and incompleteness of some of Koppitz's scoring instructions were more than minor inconveniences. They had three major consequences: (1) inter-scorer reliability between judges not given the opportunity to discuss scoring interpretations was adversely affected, (2) exact replication of the present study will be difficult since future researchers may make scoring interpretations quite different from those made by E, and (3) results of the

present study reflect to some extent the reliability of scoring interpretations made by E as well as the reliability of Koppitz's scoring system.

From the positive side, it is fair to say that the majority of Koppitz's scoring instructions can be easily and objectively followed.

Implications of the Results

The presence of two or more Emotional Indicators on the HFD of a child is, in Koppitz's words, "highly suggestive of emotional problems and unsatisfactory interpersonal relationships" (1968, p. 42). If the findings of future studies agree with the results of the present study, this two-or-more diagnostic category will be found too unreliable for clinical use.

Much the same is true for the individual stability of the majority of Koppitz's Emotional Indicators. Phi coefficients too small to be considered of predictive or practical value were found for 28 of her 30 signs.

On the other hand, the results of the present investigation indicated that reasonably good inter-scorer consistency can be attained using Koppitz's scoring system, provided judges have a chance to discuss and agree upon common scoring interpretations before scoring is begun.

The results of the present investigation are suggestive, but inconclusive. Firm conclusions about the reliability of a test or scoring system cannot be drawn from the

results of a single study, especially not from one employing a narrow age range of Ss as was the case in the present investigation. Further study is needed.

Suggestions for Further Research

Roback (1966), after a lengthy review of HFD research, concluded, "In addition to the paucity of quality research in this area, it is obvious that there is a great need for standardized and validated scales for estimating personality adjustment from figure drawings" (p. 16). From this point of view, Koppitz's attempt at a new, objective scoring system for HFDs is a step in the right direction. However, as mentioned earlier, some of Koppitz's scoring directions are difficult to interpret and apply. It is suggested that future researchers seek clarification of these scoring instructions before pursuing studies of the system's reliability.

Results of the present study were limited in scope since the majority of the Ss used were eight and nine years of age. It is recommended that future studies employ Ss from all age levels covered by Koppitz's system.

Most importantly, appropriate statistical treatment of results should be carefully considered by future researchers. Much of the previous work concerned with the reliability of HFD signs is difficult to interpret since most authors have not reported reliability findings in terms of degree.

In the 1967 revision of the Publication Manual of the American Psychological Association, this recommendation was made.

> When a study reports the predictive validity of a method of measurement, the results must always show the degree of relationship between the measure and the criterion, and its practical value. The relationship should be reported in such terms as coefficients of correlation, cost-utility data, or expectancy tables. It is not sufficient merely to show that the relationship is nonchance in terms of the level of significance (p. 13).

This recommendation would be well applied to future studies of reliability. Correlation coefficients seem the most appropriate for reporting the reliability of HFD signs, although other measures of degree may be equally valid.

The percentage of agreement between scores or scorers does not qualify as a valid measure of relationship degree. Even when coupled with significance levels, it is misleading and should not be employed. This point can be illustrated from the results of the present study. The degree of the test-retest relationship, as measured by phi, for the Emotional Indicator clouds was found to be .19. This finding accurately indicates low test-retest stability.

However, using the same data, from the same study, for the same sign, the percentage of agreement between test-retest scores is found to equal .95, significant at the .01 level. Such a finding, presented without explanation, would probably be misinterpreted by most observers as indicating high reliability. For this reason, use of the statistic, percentage of agreement, is not recommended.

CHAPTER V

SUMMARY

In 1968, Elizabeth Koppitz introduced a new "Emotional Indicator" scoring system for the HFDs of children age five to twelve years. Precise reliability information was not included. The present study attempted to gather and evaluate reliability data for this new scoring scheme.

Test-retest HFDs were secured from 438 Moses Lake, Washington, third grade pupils. A test-retest interval of one month was employed. All HFDs were scored by E. Additional judges scored 50 of the same drawings.

Phi coefficients were calculated to determine test-retest stability and inter-scorer consistency. Agreement between scorers given the opportunity to discuss instructions prior to scoring was found to be reasonably good. Low test-retest reliability was found for the majority of Koppitz's Emotional Indicators and for her diagnostic scoring category, two or more signs present.

Clarification of Koppitz's scoring instructions and further study were recommended.

REFERENCES

American Psychological Association, Council of Editors. Publication Manual of the American Psychological Association. (Rev. ed.) Washington, D. C.: American Psychological Association, 1967.

Apfeldorf, M., Randolph, J. J., & Whitman, G. L. Figure drawing correlates of furlough utilization in an aged population. Journal of Projective Techniques, 1966, 30, 467-470.

Bradshaw, D. H. A study of group consistencies on the Draw-A-Person Test in relation to personality projection. Unpublished master's thesis, Catholic University, 1952. Cited by C. H. Swensen, Empirical evaluations of human figure drawings. Psychological Bulletin, 1957, 54, 431-435.

Craddick, R. A., Leipold, W. D., & Cacavas, P. D. The relationship of shading on the Draw-A-Person Test to Manifest Anxiety scores. Journal of Consulting Psychology, 1962, 25, 193.

French, J. W., & Michael, W. B. Standards for Educational and Psychological Tests and Manuals. Washington D. C.: American Psychological Association, Inc., 1966.

Garrett, H. E. Statistics in Psychology and Education. (5th ed.) New York: David McKay, 1958.

Grams, A., & Pinder, L. Signs of homosexuality in human figure drawings. Journal of Consulting Psychology, 1958 22, 394.

Guilford, J. P. Fundamental Statistics in Psychology and Education. (3rd ed.) New York: McGraw-Hill, 1956.

Guinan, J. F., & Hurley, J. P. An investigation of the reliability of human figure drawings. Journal of Projective Techniques, 1965, 29, 300-304.

Hall, L. P., & Ladriere, L. A comparative study of diagnostic potential and efficiency of six scoring systems applied to children's figure drawings, Psychology in the Schools, 1970, 7, 244-247.

Hammer, M., & Kaplan, A. M. Reliability of profile and front-facing directions in children's drawings. Child Development, 1964, 35, 973-977. (a)

Hammer, M., & Kaplan, A. M. The reliability of sex of first figure drawn by children. Journal of Clinical Psychology, 1964, 20, 251-252. (b)

Hammer, M., & Kaplan, A. M. The reliability of size of children's drawings. Journal of Clinical Psychology, 1964, 20, 121. (c)

Hammer, M., & Kaplan, A. M. The reliability of children's human figure drawings. Journal of Clinical Psychology, 1966, 22, 316-319.

Handler, L., & Reyher, J. The effects of stress on the Draw-A-Person Test. Journal of Consulting Psychology, 1964, 28, 259-264.

Hoyt, T. E., & Baron, M. R. Anxiety indices in same-sex drawings of psychiatric patients with high and low MAS scores. Journal of Consulting Psychology, 1958, 23, 448-452.

Koppitz, E. M. Psychological Evaluation of Children's Human Figure Drawings. New York: Grune and Stratton, 1968.

Lehner, G. F. J., & Gunderson, E. K. Reliability of graphic indices in a projective test (Draw-A-Person). Journal of Clinical Psychology, 1952, 8, 125-128.

Litt, S., & Margoshes, A. Sex change in successive Draw-A-Man-Tests. Journal of Clinical Psychology, 1966, 22, 471.

Mogar, R. Anxiety indices in human figure drawings. Journal of Consulting Psychology, 1962, 26, 108.

Nichols, R. C., & Strumpfer, D. J. W. A factor analysis of Draw-A-Person Test scores. Journal of Consulting Psychology, 1962, 26, 156-161.

Roback, H. B. Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. Psychological Bulletin, 1968, 65, 1-19.

Snedecor, G. W., & Cochran, W. G. Statistical Methods. (6th ed.) Ames, Iowa: The Iowa State University Press, 1967.

Solar, D., Bruehl, D., & Kovacs, J. The Draw-A-Person Test: Social conformity or artistic ability? Journal of Clinical Psychology, 1970, 26, 524-525.

Sopchak, A. L. Anxiety indicators on the Draw-A-Person Test for clinic and nonclinic boys and their parents. The Journal of Psychology, 1970, 76, 251-260.

Starr, S., & Marcuse, F. L. Reliability in the 'Draw-A-Person' Test. Journal of Projective Techniques, 1959, 23, 83-86.

Strumpfer, D. J. W. The relation of Draw-A-Person Test variables to age and chronicity in psychotic groups. Journal of Clinical Psychology, 1963, 19, 208-211.

Swensen, C. H. Empirical evaluations of human figure drawings. Psychological Bulletin, 1957, 54, 431-466.

Swensen, C. H. Empirical evaluations of human figure drawings: 1957-1966. Psychological Bulletin, 1968, 65, 20-44.

APPENDIX A

APPENDIX A

LETTER OF EXPLANATION FOR TEACHERS

Dear Mrs. _____:

Mrs. Marlene Jensen, Midway teacher aide, will be in your room on _____, from _____ to give your students a brief test. Mrs. Jensen will handle the testing but your presence will be necessary since she is not a certificated teacher.

This testing is part of a study I am conducting for my Master's Degree thesis. I am attempting to determine the reliability of a new scoring system for the Draw-A-Man Test. The results will indirectly benefit all school districts, including Moses Lake, where this test is now being used. Your cooperation will help make this possible. You will be sent a summary of all findings as soon as the study is completed.

Since it will be necessary to repeat the test during the latter part of May, please do not discuss the testing with your pupils or have them practice drawing human figures. They should not know that they will be retested in May.

To assure anonymity and thereby confidentiality, students' names will be removed from their drawings before any analysis of the collected data is attempted.

If you have any questions or anticipate any difficulties, please contact your building principal or give me a call. Thank you.

Mark Miles
School Psychologist

APPENDIX B

APPENDIX B

TEST PROCEDURE

1. SAY: We are going to do something special today. Clear off your desks. Put away all papers, pencils, and books. You will need to listen carefully so please, no talking.

2. Position desk dividers.

3. SAY: Don't touch the paper and pencil I am going to give you until I tell you to do so. Don't touch them.

4. Pass out paper and pencils.

5. SAY: Write your name in the top right hand corner of the paper. (demonstrate) Put both your first and last name. When you have finished writing your name, put your pencil down and listen.

6. SAY: Do not draw until I tell you to begin. On this piece of paper, I would like you to draw a WHOLE person. It can be any kind of person you want to draw. Just make sure that it is a whole person and not a stick figure or a cartoon figure. You may draw a man or a woman or a boy or a girl, whichever you want to draw. Remember, no talking and do your own work. When you are finished or if you have a question, raise your hand. Now, begin.

7. Begin timing.

8. If a student raises his hand, motion him to the front of the room and answer his question privately.

9. Collect drawings as they are completed. Ask children who have finished to put their heads down.

10. After 15 minutes collect all remaining drawings. Mark drawings which children are still working on with an X.

Note: While children are drawing, have classroom teacher list the names of all children who draw with their left hand.

APPENDIX C

APPENDIX C

TEST-RETEST SCORES

Emotional Indicator	N	Pre-A* Pre-M	Pre-A Abs-M	Abs-A Pre-M	Abs-A Abs-M
1. Poor integration**	438	50	69	47	272
2. Shading face	438	10	28	45	355
3. Shading body	384	56	61	78	189
4. Shading hands	437	7	24	30	376
5. Asymmetry	438	6	31	38	363
6. Slanting figure	438	23	34	50	331
7. Tiny figure	438	6	20	14	398
8. Big figure	435	10	18	9	398
9. Transparency	438	8	13	19	398
10. Tiny head	438	0	2	3	433
11. Crossed eyes	438	3	1	9	425
12. Teeth	438	11	25	31	371
13. Short arms	438	16	39	24	359
14. Long arms	438	0	7	11	420
15. Clinging arms	438	12	19	19	388

* Pre stands for present, Abs stands for absent, A stands for April test, M stands for May test.

** Example: Poor integration was present on both HFDs of 50 Ss, present on the April HFD but absent from the May HFD of 69 Ss, absent from the April HFD but present on the May HFD of 47 Ss, and absent from both HFDs of 272 Ss.

TEST-RETEST SCORES (Continued)

Emotional Indicator	N	Pre-A* Pre-M	Pre-A Abs-M	Abs-A Pre-M	Abs-A Abs-M
16. Big hands	438	0	3	1	434
17. Hands cut off	438	4	11	26	397
18. Legs together	438	7	17	12	402
19. Genitals	438	2	14	10	412
20. Monster	438	3	0	3	432
21. Three figures	438	0	0	2	436
22. Clouds	438	2	4	14	418
23. No eyes	438	0	4	2	432
24. No nose	438	8	10	18	402
25. No mouth	438	0	8	4	426
26. No body	438	1	1	0	436
27. No arms	438	0	1	3	434
28. No legs	438	0	1	2	435
29. No feet	386	7	17	16	346
30. No neck	185	6	11	10	158
Total Score-- two or more signs present	438	156	74	97	111

* Pre stands for present, Abs stands for absent, a stands for April test, M stands for May test.

APPENDIX D

APPENDIX D

INTER-SCORER RELIABILITY SCORES ON FIRST* 25 HFDS

Code number of HFD	E	Teacher	Psychologist
57430	1,19**	1,2	3
69296	1,5	1,2,6	5,8,9,10,17
83048	3,8,13	8,13	8,13
22626	1,3	2	1,10,13,30
24480	6,8,15,30	6,8,15	4,8,15, 20,25,30
34936	--	1,6,13	3,13,30
39936	2	2	13
83671	--	2,22	13,22
80827	3,8	2	8
96363	1,3,13,23,25	1,2,3, 13,23,25	1,3,5, 13,23,25,30
35213	1	1,6	13,30
39495	3	2,3	--
89755	15,24	2,15,24	15,24
20632	3,18	2,18	13
35067	--	2	2,13,30

* These HFDs were scored prior to training of judges by E.

** Emotional Indicators are numbered from 1 to 30 (see page 1). For example, E found the signs, poor integration and genitals present on HFD #57430.

INTER-SCORER RELIABILITY SCORES ON FIRST* 25 HFDS (Continued)

Code number of HFD	E	Teacher	Psychologist
43093	1,2,3,4	1,2,3, 6,13,15,30	3,13,15,30
66598	3,4,7	2,3,4,7,24	3,4,7
23746	13,29	13	13,29,30
29226	1,6,13,26	1,2,6,13,26	1,5, 13,20,26,28
76923=	1,9,12,30	1,2,9,12,30	9,12,20,30
50803	3	2	10
71153	14	1,5	5,14
57194	3,8	3	8
33851	2,3,13,14	2,3,4	2,3,4,5,13
09419	--	1,2,6	--

* These HFDs were scored prior to training of judges by E.

APPENDIX E

APPENDIX E

INTER-SCORER RELIABILITY SCORES ON SECOND* 25 HFDS

Code number of HFD	E	Teacher	Psychologist
00042	--	3	3
77888	30**	30	1,30
34421	1,5,12	2,12	3,5,12,14
20384	3	3	3
76128	--	--	--
68243	3,11	3	3
79000	1,15,30	1,15	1,15,30
73598	2,3,4,12	2,3,4,12	2,3,4,12
12691	3,25	3,25	3,4,25
55108	3	2,3	3,4
73579	3,15	1,15	1,3,15
46703	3	3	18,19
23608	1,3,4,11	1,3,4,11	1,3,4,11
16269	8,15	3,8,15	8,15
96325	1,6,12,15,30	12,15,30	1,6,12,15,30
52936	1,3,7	1,3,7	3,7

* These HFDs were scored following training of judges by E.

** Emotional Indicators are numbered from 1 to 30 (see page 1). For example, E found the sign, no neck, present on HFD #77888.

INTER-SCORER RELIABILITY SCORES ON SECOND* 25 HFDS (Cont'd.)

Code number of HFD	E	Teacher	Psychologist
51117	1	1	1
89950	1,2	1,2	1,2,3
24969	8	1,8	1,8
71659	3,6,13	3,6	6
85651	6	--	6
16916	1	1	1
32847	1,9	1,2,9	2,9
92973	--	--	11
11591	1,3	1,2,3	1,3