1965

Analysis of Examiner Variance on the Wechsler Adult Intelligence Scale

John Joseph Henning
Loyola University Chicago

ANALYSIS OF EXAMINER VARIANCE ON THE WECHSLER ADULT INTELLIGENCE SCALE

by

John J. Henning

A Thesis submitted to the Faculty of the Graduate School of Loyola University in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

June

1965

Life

John Joseph Henning was born in Chicago, Illinois, June 18, 1939. He was graduated from Mt. Carmel High School, Chicago, Illinois, June 1957, and from Loyola University, Chicago, Illinois in February 1962, with the degree of Bachelor of Science.

He served as graduate assistant in the Psychology Department of Loyola University for the academic years 1962-1963 and 1963-1964. In June, 1964 he began serving his clinical psychology clerkship at the Illinois Youth Commission, Reception and Diagnostic Center for juvenile delinquents. He has continued at the Reception and Diagnostic Center with clinical diagnostic and clinical research duties.

Acknowledgments

The writer is indebted to Paul J. von Ebers Ph.D., Assistant Professor, Loyola University, under whose direction this study was completed. Acknowledgment is also made to Russell H. Levy Ph.D., Director of Clinical Services at the Reception and Diagnostic Center for his assistance and encouragement throughout this project. The writer also wishes to express his appreciation to Patrick M. Henning for technical help with computor processing of the data and Miss Janet A. Brehm for her assistance in the preparation of this manuscript.

Table of Contents

Chapter		Page
I.	Introduction	1
II.	Related Literature	7
III.	Design of Research	16
	A. Selection of the Test Records	18
	B. Statistical Treatment of the data	22
IV.	Results	23
V.	Summary and Conclusions	26
Bibliography		29
Appendix		33

List of Tables

Table		Page
I.	Significance of Examiner Variance on Full Scale I.Q., Verbal I.Q., Performance I.Q., and Each of the Subtests of the WAIS	34
II.	Analysis of WAIS Full Scale I.Q. Examiner Variance	35
III.	Analysis of WAIS Verbal I.Q. Examiner Variance	36
IV.	Analysis of WAIS Performance I.Q. Examiner Variance	37
V.	Analysis of WAIS Information Examiner Variance	38
VI.	Analysis of WAIS Comprehension Examiner Variance	39
VII.	Analysis of WAIS Arithmetic Examiner Variance	40
VIII.	Analysis of WAIS Similarities Examiner Variance	41

IX. Analysis of WAIS Digit Span Examiner Variance 42

X. Analysis of WAIS Vocabulary Examiner Variance 43

XI. Analysis of WAIS Digit Symbol Examiner Variance 44

XII. Analysis of WAIS Picture Completion Examiner Variance 45

XIII. Analysis of WAIS Block Design Examiner Variance 46

XIV. Analysis of WAIS Picture Arrangement Examiner Variance 47

XV. Analysis of WAIS Object Assembly Examiner Variance 48

Chapter I

Introduction

The Wechsler-Bellevue Intelligence Scale published in 1939 had become second only to the Stanford-Binet Intelligence Scale in frequency of administration in less than ten years (Loutitt and Browne 1947). A wealth of research on the Wechsler-Bellevue suggested many areas of possible improvement.

> "The chief weakness of the Wechsler-Bellevue stemmed from the unrepresentativeness of its normative sample, which was drawn largely from New York City and its environs. The total number of adults of both sexes included in this sample was only 1081. The reliability of some of the subtests was quite low, especially for proposed profile analysis of subtest scores. Obsolescent items, meager validity data, and inadequacies of the manual were among the other deficiencies of this scale" (Anastasi 1961, p. 304).

In 1955 the Wechsler Adult Intelligence Scale, hereafter referred to as the WAIS, was issued as a revised and restandardized Wechsler-Bellevue. The WAIS, therefore, emerged as a second generation Wechsler test that profited from the wealth of research and criticism on the first generation instrument. Now in its turn the WAIS has become the focus of

renewed research that may point the way to further improvements in the test with the possible production of a third generation Wechsler incorporating these suggestions. The indication that the Wechsler instrument is an evolutionary creature sensitive to its environment and capable of adaptation portends a long and robust life for the test. The widespread use of the WAIS is indicated by the purchase of between 100,000 and 1,000,000 WAIS test forms in 1964 alone (The Psychological Corporation, 1965). Because of its extensive use, its adaptability, a decade of active existence, and projected longevity, the WAIS seems one test on which criticism may have a practical rather than wholly theoretical value, and therefore, was chosen as the focus of this thesis.

The WAIS has a two-fold importance to the clinical psychologist as stated by Wechsler (1958, p. 155),

> "Although the primary purpose of an intelligence examination is to give a valid and reliable measure of the subject's global intellectual capacity, it is reasonable to expect that any well conceived intelligence scale will furnish its user with something more than an I.Q. or M.A. .... data regarding the testee's mode of reaction, his special abilities or disabilities and, not infrequently, some indication of his personality traits."

And in fact, Wechsler (1958) does suggest a "method of successive sieves" to arrive at a psychological diagnosis on the basis of WAIS part and whole patterns. While many authors have also suggested patterns for psychological diagnosis from the WAIS, Jones (1956) and McNemar (1957) maintain that the search is one in vain because of the inherent characteristics of the subtest reliabilities.

Whether the clinical psychologist is primarily interested in the psychometric I.Q. or in the psychodynamics revealed by intra-person subtest variations, the reliability of the WAIS as a measuring instrument is assumed. Wechsler (1955 p. 13) cautions, " ... the lower the reliability of the scores, the more likelihood there is that the differences between them is due to chance rather than to any real difference in the abilities possessed by the subject." The WAIS manual does present reliability coefficients for Full Scale I.Q., Verbal I.Q., Performance I.Q., and each of the subtests. In view of Wechsler's admonition to interpret subtest pattern psychograms in relation to subtest reliabilities, a caution which is repeated on each WAIS test form, it is of considerable importance that the types of reliability estimates be accurate and adequate. Cronbach and Azuma (1962) have discussed extrinsic factors which affect measures of

internal reliability such as the split-half reliability coefficient. Jastak and Jastak (1964) mention both assumed credit for items at the lower end of the WAIS Vocabulary subtest and assumed failures at the upper end of that subtest as contributing to unreliability. This criticism can be extended to other subtests with a similar characteristic: Information, Comprehension, and Arithmetic. And Similarities, Digit Span, and Block Design have assumed upper end failures only. The procedure of assigning scores for items not administered can only lead to an artificial rise in the split-half reliability coefficients.

Anastasi (1961 p. 111) indicates the need for reliability estimates other than those usually considered by the author of a test. "In tests in which examiner idiosyncrasy may play an appreciable part, it appears desirable to obtain some measure of the 'examiner reliability' of the test, especially when results by several examiners are to be combined." The standardization information for the WAIS was obtained by combining the test results from "some 77 trained examiners" (Wechsler 1955).

Anastasi (1961 p. 111) also suggests another type of reliability estimate that is applicable to individually administered and scored tests such as the WAIS. "Many current projective techniques leave much to the subjective interpretation

of the scorer, who is also usually the examiner .... For such tests, there appears to be fully as much need for an index of scorer reliability as for the more usual measures of reliability."

The examiner variance, reliability, error, or bias investigated in this thesis is a composite of several sources of error. Whether the examiner introduces an error in addition to that indicated by the split-half reliability coefficient or whether the examiner error is included in the unexplained variance is unknown. The scorer error is one element of examiner bias, since some of the subtests of the WAIS require judgment on the part of the examiner. Concerning the WAIS Wechsler (1955 p. 29) states: "For all of the Performance tests and three of the Verbal tests, the scoring is completely objective. However, evaluation of responses in the Comprehension, Similarities and Vocabulary tests demands considerable judgment by the examiner."

Clinical interaction is a global term used to describe that portion of the examiner error which is the result of the subject's personality characteristics, the examiner's personality characteristics, the physical environment in which both examiner and subject find themselves at the time of testing and the interaction of all three. Implied by Wechsler (1955 p. 26)

in his emphasis on standard procedures is an awareness of the variables of physical environment: "The room where the testing is done should be free from distracting noises and intrusions .... well lighted and ventilated .... furniture .... so arranged that the subject and examiner are comfortable, the subject can manipulate the performance materials freely, and the examiner can present the materials .... conveniently." Also implicit in these instructions is a limited attempt to standardize the subject-examiner interaction. " .... ostentatious concealment of the materials may elicit an unfavorable reaction from the subject." Sufficient time should be scheduled for the testing so that good rapport may be established and maintained and that the administration may proceed in an easy, unhurried manner." Cooperation, motivation and encouragement are also cited as necessary for proper testing (Wechsler 1955 p. 27). To "standardize" the examiner training in individual testing and special training in the WAIS is necessary according to Wechsler (1955 p. 26).

Chapter II

Related Literature

A. Research on Examiner Bias in the Wechsler Scales

Cohen (1950) investigated 13 sets of records from clinical psychology trainees who administered the Wechsler-Bellevue, the form of which was not specified. Each examiner had between 17 and 35 protocols. A test of the per cent average subtest contribution to the total weighted Wechsler-Bellevue score revealed examiner bias on only one subtest, Arithmetic, and for only one examiner. Examiner bias was, however, suggested to Cohen by the rank order correlation coefficient between smallness of inter-examiner variation on a subtest and validity of the subtest. The correlation coefficient of $.59^{\pm}$ .21 was obtained.

Masling (1959) found that the scoring of Information, Comprehension, and Similarities subtests of the Wechsler-Bellevue Form II were influenced by the attitude of the testee. Graduate students with at least one course in individual intelligence test administration gave and scored test proto-

cols given by subjects who were confederates of the author and who played a prescribed "warm" or "cold" role in the testing situation. The warm role was one in which the testee acted approving and interested, while the cold role cast the testee as rejecting and disinteredted in the test. Ten examiners turned in usable tests on one each of a cold and warm subject. The examiner was unaware that his subjects were playing roles. Even though the subjects recited a memorized list of answers for the cold and warm condition leniency in scoring significantly favored the "warm" testees. Eight of the ten examiners recorded each of his test sessions. An analysis of the test situation indicated that more reinforcement comments and more opportunities for clarification of answers were given to the warm group than the cold group.

Walker (1964) reports scoring difficult in a study of the WAIS Comprehension items. Two clinicians independently scored the same set of 500 Comprehension items taken from 50 protocals administered by the same testees. There was 78% agreement on these items. The 22% of the items on which the two clinicians disagreed were then submitted independently to five ABEPP diplomatesfor scoring. These clinicians showed unanimous agreement on only 24% of the items submitted to them.

If one assumes that the diplomateswould also have agreed with the judgment of the first two clinicians, then the total agreement on 500 items would be 81%. The Comprehension subtest therefore, seems highly vulnerable to examiner error because of the subjectivity of the scoring. This study is in agreement with a previous indication that neither clinical students nor clinicians showed high interscorer reliability on the Comprehension portion of the Wechsler-Bellevue Form I (Plumb and Charles 1955).

Murdy (1962) analyzed examiner variance on the WAIS Full Scale Scores obtained from 48 male students in a general psychology course. A group paper and pencil test of intelligence was administered to selected students. Percentile rankings that would give 48 students an I.Q. distribution similar to a normal distribution were included in the experiment. Eight male graduate students who had completed a course in individual intelligence testing were the examiners. Each subject was given the Information, Similarities, Vocabulary, Picture Arrangement, and Object Assembly subtests from one examiner, and the Comprehension, Arithmetic, Digit Symbol, Picture Completion and Block Design subtests from a second examiner. Examiners were alternately positive, warm, approving,

and interesting for one set of subtests and negative, rejecting, and disinterested when administering another set of subtests. All subjects were exposed to half the test in each manner and in this way served as their own control. Counterbalanced pentad presentation and positive and negative administration were obtained. The examiners were paired off and each tested twelve subjects equated in intelligence according to a group test. Because the two pentads used had been shown by previous studies to correlate in the 90's with the WAIS Full Scale Score, no real difference was expected between the scores obtained by the two pentads unless the difference resulted from the examiners on the treatments. No significant score differences were found to have resulted from the use of four pairs of examiners or from positive and negative treatments. Only on the Vocabulary subtest scores did a mean difference between positive and negative administration occur at the two percent level.

B. Research on Examiner Bias in the Binet Scales

Of the 60 examiners who took part in the Harvard Growth Study, 25 examiners gave Stanford-Binet intelligence tests to pupils for whom a second Stanford-Binet score was available. Cattell (1937) compared the median and upper and lower quartile

point differences between the scores obtained by one examiner and the other available score. A visual analysis of Cattell's graphed medians and quartile ranges indicates both inter-examiner median and range differences. No statistical analysis of the data was performed.

Gordon and Durea (1948) tested a group of students with the Revised Stanford-Binet Form L and retested the same group two weeks later with comparable items of the Revised Stanford-Binet Form M. Before the second test discouragement was introduced to one half of the original group through failure on some tasks. An analysis of covariance was used to test the mean scores of the discouraged and control groups in order that the unequal group mean I.Q.'s could be adjusted. Discouragement lowered performance. Tiber and Kennedy (1964) gave incentives of verbal praise, verbal reproof, candy rewards or standard administration of the 1960 Stanford-Binet Form L-M without rewards to each of four groups of children. None of these groups scored significantly better on the Binet than the control group. Klugman (1944) produces better test scores for white children compared with Negro children on the Revised Stanford-Binet with verbal praise; Negro children, however, showed more test score improvement than white children when

the reward was monetary.

C. Research on Examiner Bias in Projectives and Other Tests

Wickes (1956) examined the effect of perfunctory, verbal comments, and perfunctory non-verbal actions of smiling, nodding, and leaning forward. Groups of six college students were tested on each one of the three conditions where the examiner administered reinforcing verbal comments after every M production on Rorschach-like cards, perfunctory non-verbal reinforcement on M responses or no reinforcement. Both types of reinforcement produced significantly more M production than was found in the control group. No significant difference was found between the mean of M's produced by the two examiners.

A U. S. Army Air Force Aviation Psychology Program Research Report (cited in Lord 1950 p. 2) showed that the total number of responses given on one set of Rorschach records was a function of the examiner. The nine test examiners studied were combined in all possible pairs. A t test on the 36 pairs yielded 12 means significantly different at the 1 per cent level and three mean response differences significant at the 5 per cent level.

Lord (1950) studied the variation in the number, content, location, and determinants of Rorschach responses as a function of three separate examiners. Three examiners gave 12 tests each to the same 36 male college sophomores. The orders in which examiners administered the tests were counterbalanced. More significant t tests were related to examiner differences than to differences caused by "accepting" or "rejecting" roles played by the examiners. Affective roles produced significant response differences in thirteen Rorschach functions while examiner differences accounted for twenty-seven response category differences.

An analysis of variance performed on T.A.T. stories rated for emotional tone, outcome, and level of response indicated that the presence of an examiner while the stories were written or spoken inhibited the responses as compared with stories told or written while the examiner was absent (Bernstein 1956).

The examiner has been demonstrated to influence even the results of visual acuity tests. In an analysis of 14 different eye charts three of these charts produced significant differences in acuity scores between twelve examiners when an analysis of variance was applied to the data (U. S. Department

of the Army 1948). Two of these charts produce examiner error significant at the one per cent level, while the other chart showed examiner variance at the five per cent level.

While the literature on examiner error is not abundant it does suggest the existence of such an error. Cattell (1937) gave a clear visual picture of Stanford-Binet examiner error but did not supply any statistical considerations of his data. Since tests which appear to be more objective than the WAIS in administration and scoring have produced examiner error (U. S. Department of the Army 1948) it would seem reasonable to expect this type of error on the more subjective instrument. Cohen (1950) found this to be true only on the Wechsler-Bellevue Arithmetic subtest. Masling (1959) demonstrated examiner bias on the Information, Comprehension, and Similarities subtest as a result of the attitude that the testee assumes. Walker (1964) and Plumb and Charles (1955) observed inter-scorer variations on the WAIS and Wechsler-Bellevue Form I respectively. Murdy (1962) produced no significant examiner error in the Full Scale WAIS Scores, nor did his study produce subtest score differences between positive and negative administration with the exception of the Vocabulary subtest. There are several possible reasons why Murdy did not find examiner bias:

1) Although the subjects in the study were preliminarily selected by a group test to produce a wide range of intellectual levels, it is unlikely that the intellectual abilities of sophomores and juniors in college differ as much as their group test scores suggest.

2) It can be anticipated that the students will be well motivated, and hence, rely less on the examiner's attitude to become task orientated.

3) Lord (1950) has shown that role playing by the examiner does not produce as much variation in responses as does a less artificial attitude.

4) Examiners who are graduate students at the same university are likely to have more uniform test administration and scoring than would be found among those educated at a variety of universities and now practicing their profession, and

5) Murdy has minimized the individual examiner differences by combining examiners into pairs for comparisons.

Chapter III

Design of Research

The 196 WAIS protocols in this study were selected from all WAIS tests given at the Reception and Diagnostic Center in Joliet from January 1963 to September 1964. The time span of these records is a function of an ongoing data processing effort on all available psychometric information from boys who have been made wards of the state of Illinois. When processing of data commenced no new protocals were added in order that the backlog of collected data might first be recorded.

The Reception and Diagnostic Center receives boys from any county within the state of Illinois when the courts label a boy delinquent and commit him to the Illinois Youth Commission. "The primary purpose of the center is to effect a comprehensive professional evaluation for each admission so as to arrive at optimum diagnostic balance between the total assets and needs of a particular youth and the most adequate programs operative within the Youth Commission,"

(Levy, Grenier, Daly, and Doran 1963 p. 1).

The boys are housed at the Center for an average stay of three weeks. Their first week includes an initial interview to gather background data, as well as fingerprinting and photographing the boys for future identification. After this process is complete the boy is assigned to one of four dormitories. He may be retained in Dormitory I, which houses proportionately boys who are a containment problem, Dormitory II, which has more aggressive boys, Dormitory III, which has a younger more passive group, or Dormitory IV, which has older boys who do not present a containment problem. The actual assignment of boys to dormitories is naturally governed by many more factors than indicated here, but this serves as a general overview.

At the end of the first week those who have been admitted to this facility for the first time receive a battery of group tests which include for those 16 years and older a Bender-Gestalt test, House-Tree-Person, Otis Test of Mental Abilities, SRA Non-Verbal test, and the Revised Beta test. The WAIS is given to the new admissions of appropriate age near the end of the boy's stay at the Reception Center.

WAIS tests used in this study were given by the psychologist in his office, which for most staff members was

located in the dormitory facilities where the boys were housed. When the number of staff psychologists exceeded the available dormitory office space, the newer psychologists were temporarily given office space in a building other than the dormitories, but a building which is on the same grounds.

All examiners used in this study are male Caucasians, who ranged from 22 to 38 years of age. Four examiners had Master's degrees in clinical psychology; one had a Master's degree in school psychology. The other two examiners were lacking two courses for a Master's degree in clinical psychology. All had had a graduate level course in individual psychological testing. All but one had a graduate level course in individual intelligence testing. Proficiency in WAIS administration had been reviewed at the time they joined the Center's staff.

A. Selection of the Test Records

WAIS records which did not have all subtest admisistered were eliminated from this study. The remaining records were grouped according to examiners. From 27 examiners and over 800 records all tests were eliminated that belonged to an examiner from whom less than 30 protocols were available.

All testees who had a school grade placement of "un-

graded" were dropped from this study. This was done to avoid the possibility that one examiner might have a sample unduly loaded with persons of suspected low intelligence.

In order to equate the subjects tested in each of the examiner groups along the variables of urban-rural and Negro-white, the records of the remaining 11 examiners were divided into groups of urban-white testees, urban-Negro testees, rural-white testees, and rural-Negro testees. "Urban" was defined as any of the 38 Illinois communities that had a population over 25,000 according to the 1960 U. S. Census Report. Because several examiners had tested one or no rural Negroes all tests in this category were removed.

To maximize both the number of examiners and the number of tests per examiner and to retain an equal number of urban whites, urban Negroes, and rural whites in each examiner group, it was necessary to select 16 urban whites, nine urban Negroes, and three rural whites per examiner. This eliminated records from all but seven examiners. The selection of the records was made by assignment of a chronological number to all tests of one examiner within the grouping under consideration. The tests for this study were then chosen from that grouping with the aid of a table of random numbers. This produced the 196 WAIS protocols; 28 tests for each of seven examiners.

Similarity of inter-examiner groups seemed essential. The four primary variables that the author seeks to demonstrate group similarity on are sex, race, urban-rural status and age. The sex ratio for both groups is the same since only males were tested. Race and urban-rural status were equated through random selection within appropriately stratified groups. The similarity of age within groups was checked with a t test of the difference between the group with the largest and the group with the smallest means. Three variables of secondary importance were also investigated for possible non-change intergroup difference with a t test of the most widely varying means; grade, number of recorded offenses, and length of residency outside of Illinois. No differences between examiner age means, grade means, offense means, or length of out-of-state residency means were significant at the five percent level. (Table I)

TABLE I

AN ANALYSIS OF AGE, GRADE, NUMBER OF OFFENSES, AND LENGTH OF OUT-OF-STATE RESIDENCE FOR THE EXAMINER GROUPS WITH THE GREATEST DISPARITY

	$MEAN_1$	$MEAN_2$	STANDARD $DEVIATION_1$	STANDARD $DEVIATION_2$	DEGREES OF FREEDOM	t VALUE
AGE (Examiners 1 & 4)	198.82	196.86	4.30	3.42	54	1.90
GRADE (Examiners 3 & 5)	9.57	9.11	1.10	.87	54	1.77
OFFENSES (Examiners 1 & 6)	3.68	5.04	2.63	3.80	54	1.53
OUT-OF-STATE RESIDENCY (Examiners 3 & 4)	1.54	3.29	2.82	5.29	54	1.75

B. Statistical Treatment of the Data

A single classification analysis of variance (McNemar 1962 p. 265) was used to determine the existence of a difference in the means of seven examiner groups for Full Scale I.Q., Verbal I.Q., Performance I.Q., and each of the 11 subtest scaled scores of the WAIS.

Normality and homogeneity of variance were not tested but assumed not to vary so markedly as to distort the analysis of variance tests. Concerning these assumptions Edwards (1960 p. 132) states the following:

> "There is considerable evidence to indicate that in the common case in experimental work where the number of observations is the same for the various treatments, the F test for the means in the analysis of variance is little influenced by heterogeneity of variance. ...since the F test is very insensitive to nonnormality and since with equal n's it is also insensitive to variance inequalities, it would be best to accept the fact that it can be used safely under most conditions."

Chapter IV

Results

The Full Scale and Verbal I.Q.'s, as well as the Comprehension, Arithmetic, Similarities, Digit Span, Digit Symbol, and Picture Completion subtests of the WAIS, reveal significant inter-examiner differences at the .05 level of confidence when an F value was computed. (Appendix - Tables I to XV). Four of the six verbal subtests yield significant examiner error while only two of the five Performance subtests indicate such a bias. The largest two F values are obtained from verbal tests of Digit Span and Comprehension. These findings suggest that the examiner error of the Full Scale I.Q. is mainly a function of examiner errors on verbal areas of the WAIS. That the verbal areas are indeed more highly loaded with low examiner reliability compared with the performance areas is supported by the minimal examiner error in the Performance I.Q.

It might have been expected that those subtests which require the greater scoring judgment by the examiner as Comprehension, Similarities, and Vocabulary items do, would show

the most examiner bias, or at least all three tests would be expected to have substantially the same examiner error. However, this is not what the data reveals. Examiner reliability for Vocabulary is better than it is for all but one performance test, while Comprehension and Similarities have poor examiner reliability. The range of possible Vocabulary scores of 0 to 80 may in large part be the stabilizing factor for this subtest (Jastak and Jastak 1964).

Information, Comprehension, Vocabulary, and Similarities, to a lesser degree, are subject to possible over-and-under-inquiry. Again no consistent pattern emerges which would lead one to suspect that the "inquiry" element has a dominant role in producing error on these four tests. Information and Vocabulary present relatively good examiner reliability, although Comprehension and Similarities do not.

Digit Symbol and Picture Completion appear to be highly objective in both scoring and administration, yet produce low examiner reliability. Likewise, Arithmetic is objective though requiring greater subject-examiner interaction. Low examiner reliability is also reported for Arithmetic.

Digit Span would seem to demand more of the examiner's

skills, to effect a proper presentation, than any of the other WAIS subtests. Grouping of numbers, inaccurate interval between digits, or failure to get the testee's attention before commencing are all possible presentation failings. These appear to be the most likely elements to account for this lowered subtest examiner reliability on the WAIS.

Chapter V

Summary and Conclusions

The purpose of this study was to determine whether any appreciable error attributable to the test examiner existed in the administration of the WAIS. From over 800 test records on boys committed to the Illinois Youth Commission Reception and Diagnostic Center, 28 WAIS protocols were selected from each of seven examiners. The test selection insured equal representation of urban-rural and Negro-white in each examiner group. The sex of the subjects presented no problem since only delinquent boys are admitted to the Reception and Diagnostic Center. Age, school grade placement, number of recorded offenses, and length of out-of-state residency were compared for subjects in each of the examiner groups. A t test on the two most extreme means for each of these four variables indicated no significant difference at the .05 level of confidence. Since each examiner group of 28 subjects was shown to be similar in sex, age, urban-rural status, Negro-white ratio, school grade placement, number of recorded offenses, and length of

out-of-state residency the groups were assumed to be equated on relevant variables so that any difference in I.Q. or subtest mean scores could be attributed to examiner differences rather than subject differences.

An analysis of variance was performed on the Full Scale I.Q., Verbal I.Q., Performance I.Q., and each of the subtests of the WAIS. Normality and homogeneity of the data was assumed, but a check was not made on the assumptions.

Significant differences were found at the .05 level between the seven examiners on Full Scale I.Q., Verbal I.Q., Comprehension, Arithmetic, Similarities, Digit Span, Digit Symbol, and Picture Completion. The author considered the examiner error on the Full Scale I.Q. to be a product of low verbal subtest examiner reliability. No pattern emerged in the F values of the 14 areas studied that suggested to the author the elements of the test administration which might account for the low examiner reliability. If further investigations of differing testee populations substiantiate the findings of this study, then the clinician must consider WAIS psychopathology patterns in the light of these results, lest he diagnose the examiner rather than the subject. And if the findings of this thesis are substantially correct the researcher will undoubtedly be unsatisfied with the gross concept of examiner relia-

bility. Future work may seek to identify and eliminate the sources of unwanted error in the WAIS which make examiner bias possible.

Bibliography

Anastasi, Anne. Psychological Testing. (2nd ed.) New York: The Macmillan Company, 1961.

Bernstein, L. The examiner as an inhibiting factor in psychological testing. J. consult. Psychol., 1956, 20, 287-290.

Cattell, P. Stanford-Binet I.Q. Variations, School and Society. 1937, 45, 615-618.

Cohen, E. Is there examiner bias on the Wechsler-Bellevue? Proceedings of the Oklahoma Academy of Science, 1950, 31, 150-153.

Cronbach, L. J. & Azuma, H. Internal Consistency Reliability formulas applied to randomly sampled single-factor tests; in empirical comparison. Educ. Psychol. Meas., 1962, 22, 645-665.

Edwards, A. E. Experimental Design in Psychological Research. New York: Holt, Rinehart and Winston, 1960.

Gordon, L. V. & Durea, M. A. The effects of discouragement on the revised Stanford-Binet scale. J. genet. Psychol., 1948, 73 201-207.

Guilford, J.P. (Ed.) Printed Classification tests. Program Research Report, Report No. 5, Washington, D. C.: Government Printing Office, 1947, 24. Clinical Type Procedures. In Lord, E. Experimentally Induced variations in Rorschach Performance.

Jastak, J. F. & Jastak, Sr. Short forms of the WAIS and WISC vocabulary subtests. J. clin. Psychol., 1964, 20, 167-199.

Jones, H. G. The evaluation of the significance of differences between scaled scores on the WAIS: the perpetration of a fallacy. J. consult. Psychol., 1956, 20, 319-320.

Klugman, S. F. The effect of money incentive versus praise upon the reliability and obtained scores of the revised Stanford-Binet test. J. gen. Psychol., 1944, 30, 255-269.

Levy, R. H., Grenier, W. J., Daly, R. M., & Doiron, R. G. Cross-Sectional Psychometric Evaluation of "Court-Labelled" Delinquent Boys. Joliet: Reception and Diagnostic Center, Illinois Youth Commission, 1963.

Lord, E. Experimentally induced variations in Rorschach Performance. Psychol. Monogr., 1950, 1-34. (Whole No. 316).

Loutitt, C. M., & Browne, C. G. The use of psychometric instruments in psychological clinics. J. consult. Psychol., 1947, 11, 49-54.

Masling, J. The effects of warm and cold interaction on the administration and scoring of an intelligence test. J. consult. Psychol., 1959, 23, 336-341.

McNemar, Q. On WAIS differences scores. J. consult. Psychol., 1957, 21, 239-240.

McNemar, Q. Psychological Statistics, (3rd Ed.) New York & London: John Wiley and Sons, Inc., 1962.

Murdy, W. G. The effect of positive and negative administration of intelligence tests. 1962.

Psychological Corporation, personal communications,1965.

Plumb, G. R. & Charles, D. C. Scoring Difficulty of Wechsler Comprehension Responses. J. Educ. Psychol. 1955, 46, 179-183.

Rosenthal, R. Experimental Attributes as Determinants of Subject's Responses. J. proj. Tech., 1963, 27, 324-331.

Tiber, N. & Kennedy, W. The effects of incentives on the intelligence test performance of different social groups. J. consult. Psychol., 1964, 28, 187.

U. S. Department of the Army, The Adjutant General's Office, Personnel Research Branch. Studies in Visual Acuity. P.R.S. Report No. 742, 1948.

Walker, R. E. Personal Communication, 1964.

Wechsler, D. Manual for the Wechsler Adult Intelligence Scale. New York: The Psychological Corporation, 1955.

Wechsler, D. The Measurement and Appraisal of Adult Intelligence. Baltimore: The Williams & Williams Company (4th Ed.) 1958.

APPENDIX

TABLE I

SIGNIFICANCE OF EXAMINER VARIANCE ON FULL SCALE I.Q., VERBAL I.Q., PERFORMANCE I.Q., AND EACH OF THE SUBTESTS OF THE WAIS

WAIS DATA	VARIANCE BETWEEN EXAMINERS	DEGREES OF FREEDOM	VARIANCE WITHIN EXAMINER GROUPS	DEGREES OF FREEDOM	F
FULL SCALE I.Q.	353.7	6	129.2	189	2.738*
VERBAL I.Q.	527.5	6	137.2	189	3.845**
PERFORMANCE I.Q.	138.7	6	139.7	189	0.993
INFORMATION	9.5	6	6.0	189	1.583
COMPREHENSION	30.5	6	7.7	189	3.961**
ARITHMETIC	18.4	6	7.4	189	2.486*
SIMILARITIES	13.1	6	6.0	189	2.183*
DIGIT SPAN	27.1	6	5.9	189	4.593**
VOCABULARY	7.7	6	5.0	189	1.540
DIGIT SYMBOL	7.9	6	3.6	189	2.194*
PICTURE COMPLETION	12.7	6	5.3	189	2.396*
BLOCK DESIGN	15.7	6	9.0	189	1.744
PICTURE ARRANGEMENT	6.9	6	4.8	189	1.438
OBJECT ASSEMBLY	27.8	6	14.9	189	1.866

* SIGNIFICANT AT THE .05 LEVEL ** SIGNIFICANT AT THE .01 LEVEL

TABLE II

ANALYSIS OF WAIS FULL SCALE I.Q. EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	97.893	15.924	3.009
EXAMINER 2	28	96.357	9.507	1.797
EXAMINER 3	28	99.179	8.998	1.700
EXAMINER 4	28	92.964	10.881	2.056
EXAMINER 5	28	93.643	8.543	1.614
EXAMINER 6	28	103.464	12.273	2.319
EXAMINER 7	28	98.250	11.711	2.213

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	26532.7	195		2.738
BETWEEN	2121.9	6	353.7	
WITHIN	24410.8	189	129.2	

TABLE III

ANALYSIS OF WAIS VERBAL I.Q. EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	98.500	17.089	3.230
EXAMINER 2	28	95.571	9.203	1.739
EXAMINER 3	28	98.571	8.421	1.592
EXAMINER 4	28	92.929	10.917	2.063
EXAMINER 5	28	92.321	8.179	1.546
EXAMINER 6	28	105.107	12.306	2.326
EXAMINER 7	28	98.786	13.237	2.502

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	29095.0	195		3.845
BETWEEN	3164.8	6	527.5	
WITHIN	25930.2	189	137.2	

TABLE IV

ANALYSIS OF WAIS PERFORMANCE I.Q. EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	97.571	14.096	2.664
EXAMINER 2	28	98.000	11.823	2.234
EXAMINER 3	28	100.071	10.252	1.937
EXAMINER 4	28	93.821	11.235	2.123
EXAMINER 5	28	97.036	11.445	2.163
EXAMINER 6	28	100.643	12.497	2.362
EXAMINER 7	28	97.786	11.010	2.081

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	27243.4	195		0.993
BETWEEN	832.4	6	138.7	
WITHIN	26411.0	189	139.7	

TABLE V

ANALYSIS OF WAIS INFORMATION EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	8.393	3.047	0.576
EXAMINER 2	28	8.036	2.117	0.400
EXAMINER 3	28	8.429	2.201	0.416
EXAMINER 4	28	7.857	2.606	0.493
EXAMINER 5	28	7.214	2.200	0.416
EXAMINER 6	28	9.000	2.194	0.415
EXAMINER 7	28	8.643	2.599	0.491

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1186.1	195		1.583
BETWEEN	57.0	6	9.5	
WITHIN	1129.1	189	6.0	

TABLE VI

ANALYSIS OF WAIS COMPREHENSION EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.393	3.938	0.744
EXAMINER 2	28	7.857	1.900	0.359
EXAMINER 3	28	8.964	2.186	0.413
EXAMINER 4	28	7.679	2.435	0.460
EXAMINER 5	28	7.964	1.666	0.315
EXAMINER 6	28	10.607	3.359	0.635
EXAMINER 7	28	9.000	3.174	0.600

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1639.6	195		3.961
BETWEEN	182.7	6	30.5	
WITHIN	1456.9	189	7.7	

TABLE VII

ANALYSIS OF WAIS ARITHMETIC EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.429	3.338	0.631
EXAMINER 2	28	9.143	2.902	0.548
EXAMINER 3	28	9.286	2.141	0.405
EXAMINER 4	28	7.964	2.380	0.450
EXAMINER 5	28	7.893	2.006	0.379
EXAMINER 6	28	10.143	2.864	0.541
EXAMINER 7	28	9.214	3.071	0.580

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1500.0	195		2.486
BETWEEN	110.2	6	18.4	
WITHIN	1389.8	189	7.4	

TABLE VIII

ANALYSIS OF WAIS SIMILARITIES EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.429	3.490	0.660
EXAMINER 2	28	9.821	2.074	0.392
EXAMINER 3	28	9.964	2.472	0.467
EXAMINER 4	28	8.679	2.278	0.430
EXAMINER 5	28	8.786	2.079	0.393
EXAMINER 6	28	10.607	1.853	0.350
EXAMINER 7	28	9.214	2.515	0.475

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1209.0	195		2.183
BETWEEN	78.8	6	13.1	
WITHIN	1130.2	189	6.0	

TABLE IX

ANALYSIS OF WAIS DIGIT SPAN EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	8.750	2.675	0.506
EXAMINER 2	28	8.500	2.575	0.487
EXAMINER 3	28	9.036	1.732	0.327
EXAMINER 4	28	8.321	2.161	0.408
EXAMINER 5	28	7.964	2.457	0.464
EXAMINER 6	28	10.893	2.393	0.452
EXAMINER 7	28	9.607	2.833	0.535

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1276.0	195		4.593
BETWEEN	162.3	6	27.1	
WITHIN	1113.7	189	5.9	

TABLE X

ANALYSIS OF WAIS VOCABULARY EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	7.786	3.201	0.605
EXAMINER 2	28	7.107	1.833	0.346
EXAMINER 3	28	7.679	1.945	0.368
EXAMINER 4	28	7.143	1.957	0.370
EXAMINER 5	28	6.750	1.669	0.315
EXAMINER 6	28	8.286	2.339	0.442
EXAMINER 7	28	7.750	2.287	0.432

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	983.0	195		1.540
BETWEEN	46.0	6	7.7	
WITHIN	937.0	189	5.0	

TABLE XI

ANALYSIS OF WAIS DIGIT SYMBOL EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	8.786	2.079	0.393
EXAMINER 2	28	8.357	2.004	0.379
EXAMINER 3	28	8.393	1.548	0.292
EXAMINER 4	28	8.000	1.540	0.291
EXAMINER 5	28	8.143	1.079	0.204
EXAMINER 6	28	9.143	2.138	0.404
EXAMINER 7	28	9.429	2.516	0.475

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	726.7	195		2.194
BETWEEN	47.2	6	7.9	
WITHIN	679.5	189	3.6	

TABLE XII

ANALYSIS OF WAIS PICTURE COMPLETION EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.750	2.351	0.444
EXAMINER 2	28	9.464	2.186	0.413
EXAMINER 3	28	10.321	2.611	0.493
EXAMINER 4	28	9.107	2.025	0.383
EXAMINER 5	28	10.000	2.419	0.457
EXAMINER 6	28	11.036	2.673	0.505
EXAMINER 7	28	9.250	1.777	0.336

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1085.4	195		2.396
BETWEEN	76.2	6	12.7	
WITHIN	1009.2	189	5.3	

TABLE XIII

ANALYSIS OF WAIS BLOCK DESIGN EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.964	3.825	0.723
EXAMINER 2	28	11.036	3.037	0.574
EXAMINER 3	28	9.929	2.734	0.517
EXAMINER 4	28	8.571	2.962	0.560
EXAMINER 5	28	9.464	3.109	0.588
EXAMINER 6	28	9.571	2.559	0.484
EXAMINER 7	28	9.357	2.642	0.499

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	1803.2	195		1.744
BETWEEN	94.1	6	15.7	
WITHIN	1709.1	189	9.0	

TABLE XIV

ANALYSIS OF WAIS PICTURE ARRANGEMENT EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.143	2.189	0.414
EXAMINER 2	28	8.286	1.922	0.363
EXAMINER 3	28	9.750	1.818	0.344
EXAMINER 4	28	9.571	2.395	0.453
EXAMINER 5	28	9.286	1.941	0.367
EXAMINER 6	28	9.643	2.231	0.422
EXAMINER 7	28	9.464	2.701	0.510

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	947.6	195		1.438
BETWEEN	41.3	6	6.9	
WITHIN	906.3	189	4.8	

TABLE XV

ANALYSIS OF WAIS OBJECT ASSEMBLY EXAMINER VARIANCE

VARIABLE	NUMBER OF TESTS	MEAN	STANDARD DEVIATION	STANDARD ERROR
EXAMINER 1	28	9.250	2.901	0.548
EXAMINER 2	28	10.071	2.892	0.547
EXAMINER 3	28	11.893	7.495	1.416
EXAMINER 4	28	8.750	2.914	0.551
EXAMINER 5	28	9.643	2.909	0.550
EXAMINER 6	28	9.857	3.027	0.572
EXAMINER 7	28	9.500	2.269	0.429

SOURCE	SUM SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
TOTAL	2980.7	195		1.866
BETWEEN	166.9	6	27.8	
WITHIN	2813.8	189	14.9	

APPROVAL SHEET

The thesis submitted by John J. Henning has been read and approved by three members of the Department of Psychology.

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

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

July 31, 1966
Date

Paul K. von Ehrenfels
Signature of Adviser (PK)