12-1970

Pretest Case History Information and Examiner Experience as Determinants of Scoring Variability on Wechsler Intelligence Tests

Joseph J. Auffrey
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses

Part of the Psychoanalysis and Psychotherapy Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2959

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
PRETEST CASE HISTORY INFORMATION
AND EXAMINER EXPERIENCE AS DETERMINANTS OF
SCORING VARIABILITY ON WECHSLER INTELLIGENCE TESTS

By

Joseph J. Auffrey, Jr.

A Thesis
Submitted to the
Faculty of the Graduate College
in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
December 1970
ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Malcolm Robertson, Dr. Bradley Huitema and Dr. Neil Kent for their guidance in preparation of this thesis. I also thank the 36 examiners who contributed valuable professional time to the completion of this study.

Joseph J. Auffrey, Jr.
AUFFREY, Jr., Joseph Joffre Laurier, 1948-
PRETEST CASE HISTORY INFORMATION AND
EXAMINER EXPERIENCE AS DETERMINANTS OF
SCORING VARIABILITY ON WECHSLER INTEL-
LIGENCE TESTS.

Western Michigan University, M.A., 1970
Psychology, clinical

University Microfilms, A XEROX Company, Ann Arbor, Michigan
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>9</td>
</tr>
<tr>
<td>Subjects</td>
<td>9</td>
</tr>
<tr>
<td>Materials</td>
<td>9</td>
</tr>
<tr>
<td>Procedure</td>
<td>13</td>
</tr>
<tr>
<td>RESULTS</td>
<td>15</td>
</tr>
<tr>
<td>Wechsler Adult Intelligence Scale</td>
<td>15</td>
</tr>
<tr>
<td>Wechsler Intelligence Scale for Children</td>
<td>19</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>23</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>28</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>30</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>32</td>
</tr>
</tbody>
</table>
INTRODUCTION

Clinical psychology, as a profession, has always emphasized assessment and psychological testing in particular, as a primary function. Recently, methodological critiques of clinical psychology have called into question the clinician's traditional role as a diagnostician. Mischel (1968) cites the need for much closer scrutiny of diagnostic methods and suggests that a typical clinical diagnosis reflects the personality of the examiner as much as that of the examinee. Meehl (1960) has emphasized the influence of a clinician's cognitive activity in making intuitive judgments, even in seemingly objective diagnostic situations.

Intelligence testing has always played a major part in clinical diagnosis, and the standardized, individual tests are particularly valued because of their seemingly infallible objectivity. Terman and Merrill (1960) have stressed the parallel of a laboratory experiment with an individual intelligence test administration, and the intelligence examiner, like the experimenter, concentrates on measuring behavior while taking precautions to minimize the error of his measurement.

The Wechsler Adult Intelligence Scale (Wechsler, 1955), hereafter referred to as WAIS, and the Wechsler Intelligence Scale for Children (Wechsler, 1949), hereafter referred to as WISC, are widely used instruments (Sundberg and Tyler, 1962) which are backed by extensive norms and impressive validity and reliability data (Guertin, Rabin, Frank and Ladd, 1962; Littel, 1960; and Buros, 1965). Although

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
standardized and specific guidelines for administration and scoring might suggest little room for examiner error, research has shown the Wechsler tests to be susceptible to several sources of scoring discrepancy. At a construct level, several of the verbal subtests (comprehension, similarities, and vocabulary) used semi-objective scoring criteria and consequently even slightly ambiguous responses become a possible source of error. Plumb and Charles (1955), Schwartz (1966) and Sattler, Winget and Roth (1969) all found significant disagreement between examiners in scoring the verbal subtests of the WAIS. Sattler, Winget and Roth (1969) and Terman (Terman and Merrill, 1960) have stressed the need for strict adherence to standardized administration procedures in an attempt to minimize total examiner error and assure the unbiased direction of that error. Nevertheless, most examiners know that although the specific instructions for administration and scoring might suggest little room for subjectivity, ambiguous responses and inconsistent subject behavior quickly destroy this expectation. Massey (as cited in Sattler, Winget and Roth, 1969) noted a significant degree of disagreement for the WISC verbal subtests when the responses were deliberately ambiguous. Massey had identical protocols scored by several school psychologists and received IQ scores between 63 – 117 for the same protocol.

Is the error occasioned by scoring unreliability in the Wechsler tests expressed as random error variance? Or does the variance become systematic in nature upon the introduction of certain variables to the test situation? Rosenthal (1966) cites vagueness and ambiguity as a source of unreliability in any situation involving examiner
observation, but reminds us that only error variance would result from such a situation. If sizeable systematic variance appears, we must search for a source of examiner bias. Rosenthal (1966) and Barber and Silver (1968) agree that an examining situation can be influenced significantly by the examiner's expectations, which, in turn, are influenced by the examiner's personality and cultural traits. Rosenthal and Jacobsen (1966) showed that a teacher's expectations about a pupil's ability levels could actually produce the expected level of responses in the pupil. Several elementary teachers were given information about the ability levels of their students, some of it false or misleading. At the end of the experimental period the children were given IQ tests by their teachers and most students received scores to fulfill the teacher's expectations, regardless of whether those expectations were well founded. Differential expectations and the accompanying bias seem to occur in the clinical situation as well. Redlich, Hollingshead and Bellis (1955) found social class differential between the client and the therapist to be a significant determinant of both the clinician's attitude toward the patient and the corresponding course of therapy. The therapists involved in this study displayed hostile attitudes towards lower class patients and typically described them as harder to work with and more bothersome than other patients. A study by Robinson and Cohen (1954) suggests that systematic individual biases also exist among clinicians. Their study of clinical psychology interns revealed discrepancies in case reporting which could only be explained by the varying perceptions of the clinician. A closer look
revealed that each clinician's perception and corresponding reports followed a personal pattern, and the authors concluded that each report reflects the personality of the examiner as well as the patient. All these studies seem to suggest that both the judgement and actions of the clinician can be affected by situational and interpersonal variables.

Several studies have explored the possibility that intelligence test results may be influenced by situational or interpersonal variables. The administration of the tests, which is cited as the crucial variable by several authors (Terman and Merrill, 1960), (Cronbach, 1960), has been a target for many researchers. Masling (1959) found that interpersonal variables, such as the warmth of the subject, affect the intelligence test administration in the same way that they affect projective test administrations. Subjects memorized ambiguous responses for the three verbal subtest of the WAIS and were instructed to act warm and friendly in some administrations and cold and aloof in others. The clinical graduate students, who were serving as examiners, not only gave the warm subjects higher mean scores but also questioned them more and reinforced their efforts more. In the same area, Murdy (1962) trained examiners to act positively or negatively towards subjects in WAIS short form administrations. He found that a majority of IQ scores decreased under the negative condition, but the magnitude of differences between the two treatments only approached significance. Sattler and Theye (1967) reviewed the literature on variables affecting individual intelligence tests and concluded that the pretest expectations of the examiner may lead him to believe that
some subjects "really know" more than others, and the amount of probing and inquiring in the administration may vary accordingly.

Sattler, Hillix, and Neher (1970) provide the only research effort to date exploring possible scoring bias on the WISC and WAIS. The authors noted that previous research on experimenter effect and halo effect has shown that anything which allows the examiner to form a hypothesis about the subject's future performance is a possible source of systematic error. They hypothesized that examiners would give more credit to responses which supposedly came from bright subjects than responses which supposedly came from dull subjects. In the first experiment, graduate students, who were just finishing a testing class, were asked to score ambiguous responses to verbal subtests after the experimenter had suggested that the responses were from bright or dull subjects. Results showed that the "bright" subjects received significantly higher vocabulary scores than the "dull" subjects, but other verbal subtests showed no significant differences. The second experiment involved the use of tape recorded administrations, some of which scored approximately 130 (superior) and others which scored approximately 90 (average). Thirteen ambiguous responses were inserted at various points in the verbal subtests of both the superior and the average recordings and the examiners were asked to score the whole tape. The results of the second experiment revealed that the "bright" subjects received more credit than the "dull" subjects for the same ambiguous responses. Apparently, the context within which they occurred had weighted the value of the ambiguous responses in a systematic fashion.
Mischel (1968, page 3) has stated that "the expectation of the investigator affects not only what he looks for but what he finds", and the research reviewed herein would seem to support this observation. Nevertheless, almost every clinical or academic institution requires intake information as an essential part of the diagnostic process. Sundberg and Tyler (1962), in a popular clinical psychology textbook, designate a prediagnostic case history briefing as a regular part of the assessment procedure. What type of expectations does the examiner form after exposure to multifaceted case history data? Will the differential expectations which might be occasioned by radically differing case history data affect the scoring of the WAIS and the WISC? It is reasonable to assume that most examiners by personal experience and/or acquaintance with normative data have formed some definite impression about which groups of people and which types of people are likely to score high or low on an intelligence test. It seems possible that these examiners, when confronted with a case history composed of pessimistic elements (in terms of group membership, personal characteristics, etc.), may form a set of expectations which differ significantly from those occasioned by an optimistic case history. In consideration of these factors it is hypothesized that examiners score the WAIS and WISC in a systematically varied fashion dependent upon the nature of pretest case history information.

Several experimenters have used examiners with varying degrees of experience while studying the reliability of intelligence test scoring. Jordan (1932) had undergraduate testing students and
experienced clinicians administer the Stanford-Binet test and concluded that no significant differences existed between the two groups. Plumb and Charles (1955) used the comprehension test from the Wechsler-Bellevue Scale and found that both experts (over 200 administrations) and novices (clinical graduate students) displayed significant scoring unreliability. The experts' scoring was remarkably similar to that of the novices and the novices proved to be no less reliable. Using the comprehension subtest from the WAIS, Schwartz (1966) showed that both experienced (average of seven years post-doctoral) and inexperienced (clinical interns) clinicians have low inter-rater reliability, and that experience of the clinician did not appear to be a critical factor in scorer unreliability. In consideration of the aforementioned past research, it is hypothesized that examiners with differing levels of experience do not differ significantly in the degree of scoring unreliability.

Summary

1. Past research suggests that the WAIS and WISC are susceptible to considerable scoring unreliability.

2. The above-mentioned unreliability may become systematic variance when certain situational and examiner variables are introduced into the situation.

3. Any examiner's expectations, regardless of how they were formed, may influence his scoring.

Hypothesis:

An examiner's scoring on the WAIS and WISC is affected by the nature of the pretest case history information. It is postulated that an optimistic case history produces higher scores on a given protocol than a corresponding pessimistic case history.

4. Research suggests that clinical and testing experience is not a
determinant of scoring unreliability. Novices score just as reliably as experts.

Hypothesis:
The amount of scoring unreliability does not differ significantly for varying levels of experience.
Subjects: Thirty-six examiners with varying levels of testing training and experience comprised the subject population for this study. Twelve subjects were selected for each of the following categories:

1. Experts
   All subjects classified as experts had at least two years of professional experience in a position which involved Wechsler testing. Since some positions cater to a clientele of a limited age group, several subjects had extensive experience with one instrument and limited experience with the other. Each subject had at least 50 WAIS administrations and 50 WISC administrations to his credit and half of the subjects claimed 500 or more administrations on at least one of the tests. Of the 12 involved, seven were engaged in "clinical" positions (i.e. mental hospital, child guidance psychologists) and the remaining five in educationally related clinical work.

2. Interns
   The 12 subjects classified as Interns were graduate students in various fields who had completed training in Wechsler testing but had limited experience. All subjects in this category had administered at least 10 Wechsler tests (either WISC or WAIS) but none had administered more than 50 WISC or 50 WAIS.

3. Novices
   Subjects for the Novice group were selected randomly from a graduate testing class in Wechsler at Western Michigan University. The class had covered the course material involving administration and scored at least one WAIS and one WISC protocol at the time of the experiment.

Materials: Two protocols, one WAIS and one WISC, were prepared by the experimenter for use in this study. Although responses for each subtest were recorded on the protocols, no scoring or scaling had been done (except for those subtests which are automatically scored as administered). The responses recorded on these protocols were wholly the inventions of the experimenter and were designed to present
an ambiguous scoring situation to the examiner. Plumb and Charles (1955) found that ambiguous, hard to score responses generally possessed any or all of the following characteristics:

1. Multiple response; when the first response to a question may be worth zero points but the second response worth two points.

2. Responses which contain words which are similar to the correct answer but slightly different in meaning.

3. Responses which are not comparable to the guideline responses listed in the manual.

The verbal responses in the experimental protocols were created using the above listed characteristics as guidelines.

Previous studies involving the Wechsler tests have used only the verbal scales as experimental variables (Plumb and Charles, 1955), (Sattler, et al., 1970), (Sattler, et al., 1969), (Masling, 1959). However in this research the experimenter wished to create a closer approximation to the clinical scoring situation, so the complete tests were used along with the official Wechsler record forms. Seventy-two Wechsler record forms were prepared, 36 identical WAIS forms and 36 identical WISC forms.

Four case history summaries, two for the WAIS and two for the WISC, were prepared by the experimenter to accompany the aforementioned protocols as independent variables. The case histories purported to describe a real person who had supposedly taken the accompanying Wechsler test.

The two case histories which were used with the WAIS were designed by the experimenter to provide a basis for differential expectations on the part of the examiner. The first case history
portrayed a bleak past with a strong tone of both cultural and ma­
terial impoverishment. Although no statements were made about intel­
lectual prowess, the expectations to be garnered from this history
could only be negative. In contrast, the second case history was
intended to convey generally optimistic impressions about the client.
Although, again, no statements were made about the abilities of the
subject, the case history, in describing a typical middle class
socialization, provided a basis for optimistic expectations on the
examiners' part. In designing the case histories, the experimenter
assumed that the expectations of testers are based, in part, on his
knowledge of the client's past (socio-economic, educational and
familial) and the group memberships and personal characteristics
associated with that past.

The case history below accompanied one-third of the WAIS proto­
cols and was created as a pessimistic history:

Name: Calvin Washington
Birthdate: 4/10/42     Age: 28     Sex: Male
Marital: Single
Nativity: Biloxi, Mississippi    Color: Black
Place of Examination: Kalamazoo State Hospital
Occupation: Laborer
Education: Fifth Grade

Calvin was born in Mississippi, the seventh of
nine children. At the age of four, his father deserted
the family and he was sent to live with relatives in
Detroit. Calvin claims to have attended a Detroit
elementary school but says he quit at the age of 13
to find a job. After leaving several jobs for various
reasons, Calvin remained unemployed until 1965 when
he was committed to the Kalamazoo State Hospital,
diagnosed as a chronic undifferentiated schizophrenic.
Calvin had experienced hallucinations for six months
prior to his commitment. Three of Calvin's siblings
are presently in state institutions for the mentally
deficient.
Several factors which might be related to intelligence test performance are immediately noticeable:

1. Race
2. Education
3. Family History
4. Marital Status

Properties related to intelligence might be inferred from other material such as birthplace, family size and job history.

The case history summary labeled optimistic, which also accompanied one-third of the WAIS protocols, is presented below:

Name: John Cabot
Birthdate: 4/10/42 Age: 28 Sex: Male
Marital: Married
Nativity: Salem, Massachusetts Color: White
Place of Examination: Home
Occupation: Salesman
Education: High School

John was born the third of three children in a well to do New England family. Although his older brothers chose the professions (ministry and law), John continued to work for his father's business and has recently opened a new branch in the Kalamazoo area. John is active in family life (two children) and in civic affairs and enjoys reading historical novels as a hobby. The subject seems extremely confident, almost brassy, in the presence of others.

John Cabot's history accounts for many of the same factors as that of Calvin Washington, but his race, his education and his family history represent assets in terms of probable intelligence, whereas Calvin's characteristics might represent liabilities. It was the experimenter's contention that these case histories differ only in their brevity from intake information which might be gathered by any clinical agency.

Two case histories, one optimistic and one pessimistic, which
accompanied the WISC record forms, are recorded in the Appendix and do not differ in concept from the two WAIS case histories cited above.

Procedure: Each WAIS and each WISC protocol was assigned to one of the following treatments:

1. pessimistic case history
2. optimistic case history
3. no case history (control)

Random assignment of case history treatments was accomplished by the use of a random numbers table. The materials were then arranged in 36 packets, each containing one WAIS and one WISC protocol along with the assigned treatment for each. A brief set of instructions accompanied each packet and asked each examiner to abide by the following:

1. Read the enclosed case history sheet before starting.
2. Use only the Wechsler manual for scoring assistance.
3. Avoid discussions of these materials with other participants until termination of the experiment.

The 36 examiners, as previously described, were selected and asked to participate. Only one examiner declined participation and an equivalent replacement was obtained. They were told only that their task was for a research project that involved the scoring of the WISC and WAIS. Each examiner was assigned a number and the corresponding packet of materials was delivered to him. They were merely asked to follow the enclosed instructions and complete the experimental task at their leisure. All materials were returned within six weeks.

In summary, this study followed a design outlined below:

1. Subjects were 36 examiners with varying levels of
training and experience.

2. Materials were WAIS protocols, one for each subject and all identical. Responses were complete for protocols but no scoring was done. Each protocol was accompanied by either a pessimistic case history, an optimistic case history or no case history. WISC protocols, one for each subject and all identical were also used. Arrangements were the same as for the WAIS.

3. Each examiner-subject was asked to score and scale one WAIS and one WISC protocol after reading the accompanying case history.
RESULTS

As results were gathered, each examiner was interviewed briefly and asked to consider the probable purpose of the study in which he had just participated. Of the Intern group and Novice group combined, several examiners ascertained that scoring was the dependent variable, but no one mentioned examiner bias or the case histories. Ten of the 12 subjects in the Expert group mentioned scoring variability among examiners as the probable dependent variable, and one subject specified examiner bias, but again no mention was made of the independent variables (level of experience and differing case histories).

Most past research dealing with variability on the Wechsler tests has utilized only the verbal parts of the test or a verbal short form (Plumb and Charles, 1955), (Schwartz, 1966), (Sattler, Winget, and Roth, 1969). As previously mentioned, the whole WISC and WAIS protocol, containing both verbal and performance subtests, was used in the present study in an effort to preserve an authentic clinical scoring situation. However, only the verbal subtests of the WISC and WAIS were compiled and analyzed for the purposes of this study.

Wechsler Adult Intelligence Scale

WAIS verbal IQ scores obtained from all groups of subjects under all treatment conditions are represented in Table I below.
Table I

WAIS verbal IQ mean scores with differing levels of scorer experience and differing case histories

<table>
<thead>
<tr>
<th>Case History</th>
<th>Pessimistic</th>
<th>Optimistic</th>
<th>No Case History</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>Experts</td>
<td>Interns</td>
<td>Novices</td>
<td>Total</td>
</tr>
<tr>
<td>Level of Experience</td>
<td>82.8</td>
<td>84.0</td>
<td>80.25</td>
<td>82.2</td>
</tr>
<tr>
<td></td>
<td>81.0</td>
<td>84.33</td>
<td>89.0</td>
<td>85.16</td>
</tr>
<tr>
<td></td>
<td>82.33</td>
<td>84.0</td>
<td>78.0</td>
<td>82.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>82.08</td>
<td>84.08</td>
<td>83.33</td>
<td>82.08</td>
</tr>
</tbody>
</table>

Each of the nine cells in Table I represents a different combination of scorer experience and case history treatment with unequal cell numbers ranging from two to six. The Expert scorers who had pessimistic case histories produced an average score of 82.8 compared to 84.0 for Interns and 80.25 for Novices with the same treatment. Experts with optimistic case histories scored a mean of 81.0 while Interns and Novices in the same treatment grouping averaged 84.33 and 89.0 respectively. For the treatment with no case history, Experts showed a mean of 82.33 in comparison to 84.0 for Interns and 78.0 for Novices. As a group, the 12 Experts tallied a mean score of 82.08 with a standard deviation of 2.18. The dozen scorers in the
Intern group averaged 84.08 with a standard deviation of 3.12 across treatments, and the 12 Novices averaged 83.33 with a standard deviation of 6.25. The pessimistic treatment group, including Experts, Interns, and Novices, received a mean score of 82.2 as compared with 85.16 for the optimistic group and 82.08 for the group which had no case history. It is apparent that variability exists between scorers of different experience and between different case histories. In order to ascertain the statistical significance of this variability and to pinpoint its origin, a close scrutiny of the data was needed.

A two factor analysis of variance design for cells containing unequal number of subjects was used to determine how much of the total variability could be accounted for by the independent variables.

Table 2

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Case History</td>
<td>17.80</td>
<td>2</td>
<td>8.90</td>
<td>1.75 not significant</td>
</tr>
<tr>
<td>Treatments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Levels of Experience</td>
<td>7.27</td>
<td>2</td>
<td>3.64</td>
<td>.71 not significant</td>
</tr>
<tr>
<td>A x B</td>
<td>51.55</td>
<td>4</td>
<td>12.88</td>
<td>2.54 not significant</td>
</tr>
<tr>
<td>Error</td>
<td>33</td>
<td></td>
<td>5.06</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 above depicts the amount of variability attributable to
each factor and provides an error term with which to gauge the significance of gross variability. The F value of 1.75 for case history treatments is not significant (alpha = .05), indicating that the scorers were not systematically biased by the nature of a pretest case history. The level of scoring experience factor also failed to significantly affect verbal IQ scores as evidenced by an F value of .71. The F ratio of 2.54 for the combined effect of experience and case history is not significant, although it should be noted that the two factors in concurrence account for more variance than either factor alone.

Table 1 above, shows that the groups representing differing levels of experience (Experts, Interns, Novices) recorded standard deviations of 2.18, 3.12 and 6.25 accordingly. Since these figures could represent differences in group scoring unreliability, Levene's test (Glass and Stanley, 1970) was used to evaluate the homogeneity of scoring variability among these groups. The test, represented in Table 3 below, is simply a one way analysis of variance of the average deviations of each group.

Table 3
Levene's test for differences in scoring variability among scorers of varying experience

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Group</td>
<td>61.01</td>
<td>2</td>
<td>30.51</td>
<td>3.29</td>
</tr>
<tr>
<td>Within Group</td>
<td>257.29</td>
<td>33</td>
<td>7.79</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>318.3</td>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The F value of 3.92 is significant (alpha = .05) and indicates that the scoring variability and hence the scoring unreliability differed significantly between scorers of different experience levels.

The statistical significance noted above tells us that a significant amount of variance occurred among the three groups of varying experience. The Tukey method (Glass and Stanley, 1970) was applied to the data in Table 3 above in an effort to determine which pairs of groups differed significantly in their scoring unreliability. Differences between the examiner groups were corrected according to the Tukey formula and evaluated by comparison with the Studentized range table. The differences between the Experts and the Interns produced a value of .432 which is not significant. The difference between Experts and Novices resulted in a value of 3.60 which is significant (alpha = .05), while the value of 3.17, representing differences between Interns and Novices, closely approaches significance (3.486 alpha = .05).

The raw data give superficial evidence of the differences found above. Verbal IQ scores received for Experts ranged from 78 to 85, all within the same classification of dull normal (Wechsler, 1955), while the Novices ranged from a borderline defective 72 to an average 96.

Wechsler Intelligence Scale for Children

WISC verbal IQ scores, obtained from all groups of scorers under all treatment conditions, are presented in Table 4 below.
<table>
<thead>
<tr>
<th>Case History</th>
<th>Pessimistic</th>
<th>Optimistic</th>
<th>No Case History</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>Experts</td>
<td>81.5</td>
<td>81.0</td>
<td>77.33</td>
</tr>
<tr>
<td></td>
<td>Interns</td>
<td>88.0</td>
<td>84.5</td>
<td>83.0</td>
</tr>
<tr>
<td></td>
<td>Novices</td>
<td>85.83</td>
<td>74.0</td>
<td>77.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>85.83</td>
<td>81.85</td>
<td>78.16</td>
</tr>
</tbody>
</table>

The table shows that Expert scorers who received pessimistic case histories produced a mean score of 81.5, while the Interns and Novices with the pessimistic treatment averaged 88.0 and 85.83 respectively. The Expert group tallied 81.0, the Interns 84.5 and the Novices 74.0 when confronted with the optimistic case history treatment. When no case history accompanied the protocol, Experts averaged 77.33 compared to 83.0 for the Interns and 77.0 for the Novices. As a group, the 12 Experts scored the verbal scales, all treatments included, to an average of 70.25 and a standard deviation of 4.23. All scorers in the Intern group averaged 85.41 with a standard deviation of 7.95, while the Novices registered a mean of 80.91 and a standard deviation of 11.91. The pessimistic treatment group received a total mean of 85.83, in contrast to the 81.85 received by the whole
optimistic group and the 78.16 representing the protocols with no accompanying case history.

Incorporating the data presented in Table 4, a two factor analysis of variance was used to determine the origin and extent of variability. Table 5 below depicts the amount of variability attributable to each factor.

Table 5
Two Factor Analysis of Variance for WISC Verbal IQ Scores

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Case History</td>
<td>64.01</td>
<td>2</td>
<td>32.0</td>
<td>1.31</td>
<td>not significant</td>
</tr>
<tr>
<td>Treatments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Levels of Experience</td>
<td>67.14</td>
<td>2</td>
<td>33.57</td>
<td>1.37</td>
<td>not significant</td>
</tr>
<tr>
<td>A x B</td>
<td>39.90</td>
<td>4</td>
<td>9.97</td>
<td>.408</td>
<td>not significant</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>23</td>
<td>24.41</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An F value of 1.31 results from the variance due to the case history factor, indicating no significant differences between treatments. In a similar fashion, the F ratios of 1.37 for levels of experience and .408 for the concurrence of both factors are not significant. The analysis indicates that despite raw score variability, the mean score differences between different levels of experience and different treatments are within the range of chance variation and scoring error.
Levene's test was used to probe possible differences in scoring unreliability among the three subject groups of varying experience. Table 6 below represents the partitioning of variance and significance tests using the average deviation scores of the different examiner groups.

Table 6
Levene's test for differences in scoring variability among scorers of varying experience

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Group</td>
<td>205.88</td>
<td>2</td>
<td>102.94</td>
<td>3.198</td>
</tr>
<tr>
<td>Within Group</td>
<td>1062.21</td>
<td>33</td>
<td>32.18</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1268.09</td>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As displayed by the table, the F ratio of 3.198 is close to the F value needed for significance (critical value = 3.30, alpha = .05). Thus, the seemingly substantial differences between groups may be due to chance variation rather than systematic variation.
DISCUSSION

One hypothesis offered in this study concerned the possible effect of pretest case history data on the scoring of the WAIS and the WISC. The experimenter posited that the scoring of these tests would be biased by the introduction of differential case histories and the scorer expectations which might accompany them. The results of the study were not supportive of that hypothesis. As the data in Table 1 (page 16) and Table 4 (page 20) indicates, variability was present between treatment groups, but this variability was neither consistent in direction nor large in comparison to expected error. Total scores for the WAIS suggested that the optimistic case history group might have been slightly favored in scoring, but a glance at the WISC totals showed the pessimistic group holding a favorable edge. The significance tests for the WAIS and the WISC data showed that the differential case histories did not significantly affect the scoring of the WAIS and the WISC either by themselves or in combination with differing levels of experience.

The lack of confirmation for the first hypothesis might have been due to one or all of several factors. The expected phenomena might simply have not occurred, in which case we must conclude that differential case histories have no effect upon the scoring of intelligence tests. Perhaps the experimental situation was invalid, i.e., the case histories were not representative or the scoring situation was not real enough. The significance of results may have been affected by the small samples used in this study. It also seems

23

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
possible that an examiner's awareness of the experiment, the anticipation that his moves would be counted, could have produced a scoring situation considerably different from normal. The experimenter had posited that the case histories would give rise to differential expectations on the part of the scorer and hence bias his scoring. Perhaps the experimental design of this study was too limited to detect the hypothesized phenomena. In order to comprehensively evaluate WAIS and WISC scorer bias, research involving an actual administration-scoring situation is needed. The pretest interview and the administration of the test itself might then be sources of differential expectations on the part of the scorer.

Subjects for this research included intelligence test examiners at three levels of scoring experience. As cited previously, several studies have used scorer-subjects of varying levels of experience (Jordan, 1932; Plumb and Charles, 1955; and Schwartz, 1966) and have failed to reveal significant differences in scoring agreement of unreliability. Jordan (1932) used undergraduate students in a testing class as a contrast for experienced clinicians while Plumb and Charles (1955) and Schwartz (1966) both used graduate students in clinical psychology who had previously completed testing training. The inexperienced group in the Jordan study is fairly comparable to the Novice group of the present study, while the inexperienced clinical graduate students in the latter two studies parallel the Intern group in this research. With the subject groups at least approximating those of past research, the experimenter had hypothesized a replication of previous findings that examiners of different
experience levels do not differ significantly in scoring unreliability. The results did not support the hypothesis.

As previously cited, the raw data for both WAIS and WISC showed considerable differences in the amount of variability within examiner groups. Novices produced verbal IQ scores ranging from 63 (defective range) to 106 (average range) using the same protocol, while Experts restricted their scores to a range of 71 (borderline) to 85 (dull normal). Subtest scoring range for Novices was higher in every instance than the Expert group; in one instance (comprehension subtest) the Novice scores ranged from 0 to 11 (total scoring range for test is 0 - 19). Levene's test for differences in variability confirmed what the raw data suggested; there were significant differences in reliability among the three groups in scoring the WAIS. For the WAIS, the differences closely approached the significance level. Further examination of the WAIS data indicated that the difference in unreliability between the Expert and Intern groups was insignificant, but that the Novice group was clearly separated from the other two by its lower reliability. As a whole, the data seemed to suggest that the Expert and Intern groups scored the protocols in a similar fashion or at least obtained similar scores. In contrast, the Novice group scored less reliably than the other two groups.

The factors responsible for the poor showing of the Novice group are difficult to isolate, and it seems probable that only a well controlled study of training courses and scoring procedure would yield the necessary information. As previously noted in the description of the Novice group (page 9), all had covered the training
material involved in the administration and scoring of the WAIS and WISC. The Wechsler manuals which accompany the testing materials provide "objective" scoring guidelines and general instructions for their use. If we assume that the examiners at all experience levels used the guidelines in the manual for their scoring, then it seems that the lower reliability of the Novices must be due to less skill in using the "objective" guidelines of the manual. The data for both WAIS and WISC protocols showed that scoring reliability varies in direct relation to level of experience. It would seem that the "objective" scoring guidelines of the Wechsler manuals can be used more effectively by more experienced clinicians, especially when confronted with ambiguous responses. If trained but inexperienced examiners are significantly more unreliable than those with more experience, then it seems that the needed experience should be part of the training program. Many clinical psychology and school psychology programs have recognized a need for such experience and have incorporated testing practicum as a required part of the curriculum. The trainees used in the Novice group, however, were not afforded a guided practicum as part of their training. Consequently, if introduced into a professional testing situation, they might test with less than expected accuracy and reliability. It seems that more care should be taken in the training of an examiner who must make crucial decisions on the basis of his diagnostic instruments.

In order to study more closely the experience variable in intelligence test scoring, it would seem desirable to set up a real life administration and scoring situation and to graduate the
experience variable more finely when sampling. If examiners could be subdivided into many experience levels, the nature of a relationship between experience and reliability could be explored more fully. It remains the responsibility of all clinicians to evaluate constantly the diagnostic instruments which they use and to scrutinize their own role as unbiased examiners.
SUMMARY

The Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Intelligence Scale for Children (WISC) are scored using objective scoring guidelines, but past research has shown both tests to be susceptible to scoring unreliability. Further study fails to show that the testing experience of the examiner is a significant factor in scoring reliability, but that scoring may vary systematically in accord with examiner expectations formed prior to testing. It was hypothesized that scoring of the Wechsler tests would be affected by the introduction of differing pretest case histories, and that scoring reliability for examiners of varying experience would not differ significantly.

Subjects for the study were 36 examiners at three levels of training and experience. The materials were WAIS and WISC record forms with complete but unscored responses. Each record form was accompanied by either a pessimistic case history, an optimistic case history or no case history. The case histories were deliberately designed to evoke differing expectations on the part of the examiner. Each of the 36 scorer-subjects scored and scaled one WAIS and one WISC protocol after reading the accompanying case history.

The results showed that, despite much variability, WAIS and WISC scoring were not significantly affected by the differing case history treatments. It was suggested that a more comprehensive experimental design should be used to explore the area of examiner bias.
Contrary to past research, scorer reliability was significantly affected by the training and experience level of the examiner. Statistical analysis showed that, in scoring the WAIS, Experts obtained significantly lower average deviations than the Novices. Results for the WISC showed differences closely approaching significance. The results suggested that a testing course should always include supervised practicum.
REFERENCES


30

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


APPENDIX

Case histories which accompanied WISC protocols

Optimistic

Name: Steven VanDeer  Age: 14 - 2  Sex: Male
Address: Gull Lake
Parents' Name: Dr. and Mrs. Walter VanDeer
School: Gull Lake  Grade: Ninth

Steven was born in Kalamazoo, the third of four siblings. Steven is a friendly, cheerful student according to his teachers, and his parents are well pleased with his work. All of Steven's siblings are regarded as superior students by their teachers. The home life of the VanDeer's is not only opulent but geared to the intellectual development of the youngsters. Steven has expressed the hope that he might continue his education at the University of Michigan, his father's alma mater.

Pessimistic

Name: Ricky Kowoski  Age: 14 - 2  Sex: Male
Address: 616 North Street, Kalamazoo, Michigan
Parents' Name: Victor Kowoski
School: Hurd School  Grade: Sixth

Ricky has been a constant source of trouble not only at school but in the home and on the street as well. His teachers report that he neglects his work and pesters the other students. His father reports that since his mother deserted the family two years ago, Ricky has only stopped home to eat and sleep and constantly finds trouble with the police. Three of Ricky's four siblings are behind in school, and his oldest brother attends a special education classroom. The Kalamazoo Public Schools consultant has recommended that Ricky be evaluated for possible placement at Fort Custer State Home.