# Administration Error in Presenting the WAIS-R blocks: Evaluation of an Extreme Departure

Catherine A. Bulow ©

Lakehead University

Submitted as a partial requirement
of the Master's Degree in Clinical Psychology,
at Lakehead University

ISBN 0-315-78936-0

## Acknowledgements

This thesis would not be complete without sincerely thanking my thesis advisor. I feel very fortunate to have been taught by Dr. Thompson. His indepth knowledge, counsel, criticism, and encouragement have been indispensable in the preparation of this thesis. Working with him has taught me many valuable lessons. Dr. Thompson, I thank you for your support and guidance throughout the last two years.

I would also like to extend thanks to Dr. Bauman for reviewing this thesis and offering suggestions for its improvement.

Also many thanks to David Ramsey for all his help and cooperation and to the students and staff who participated in this study.

# Table of Contents

# List of Tables

## Abstract

This study examined block presentation on the Block Design subtest of the Wechsler Adult Intelligence Scale-Revised. The standardized procedure dictates that a variety of sides be shown and that the red/white side be facing up only on one of the blocks when four are used, and only three when all nine blocks are given. It is believed that many administrators scramble the blocks, which may or may not comply with the standardized procedure. This study attempted to gauge the impact of this error on Block Design performance. Sixty subjects were tested ranging in age from 18 to 64 years and in IQ from 73 to 122. Each subject was given the Kaufman's 'quick tetrad' short form version of the WAIS-R. The Block Design subtest was administered after the tetrad, using the standardized and a nonstandardized (extreme) presentation of the blocks. The nonstandard presentation was defined as all the blocks having the full red side facing up. Results indicated the method of presenting the blocks had negligible impact on Block Design performance. There was also no significant relationship found between the short form estimate of Full Scale IQ and the difference between indices of performance for standard versus nonstandard presentations. However, these results are not conclusive, a number of pertinent factors must still be considered. It was recommended that administrators adhere to the standardized form of presentation when administering the

blocks on the Wechsler scales of intelligence.

# Introduction

Intellectual assessment is one of the cornerstones of psychological practice and has generated a large domain of theory and research. Kaufman (1990) summarized this phenomenon as follows:

> The field of intelligence, particularly of adolescent and adult mental development, has dominated the psychological literature for decades, and now encompasses a diversity of domains within cognitive psychology, clinical psychology, psychobiology, behavioral genetics, education, school psychology, sociology, neuropsychology, and everyday life (p.1).

The value of individual intelligence tests was reinforced in the United States in 1978 by Federal Law 94-142 which legitimized and confirmed their diagnostic and prognostic importance (Levenson, Golden-Scaduto, Alosa-Karpas, & Ward, 1988). In the domain of professional practice, intelligence tests are widely used and standards for their administration and application must be stringent. This study focused upon one aspect of test administration with the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981).

The Wechsler Scales of Intelligence

David Wechsler (1896-1981) developed the renowned Wechsler scales of intelligence based upon factors which he believed contributed to the overall intelligence of a person. While Wechsler regarded intelligence as being comprised of qualitatively different abilities, he did not consider it to be simply the sum of these abilities. He believed intelligent behaviour was also defined by the way abilities were combined and by a person's motivation. Wechsler (1958) described intelligence as "the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" (p.7). He therefore advocated a practical approach claiming that intelligence is known by what it enables an individual to do. Sattler (1990) noted that Wechsler published his first intelligence test, the Wechsler-Bellevue Intelligence Scale-Form 1, in 1939. A second form of this test (Form II) was published in 1946. Form 1 was revised in 1955 and became known as the Wechsler Intelligence Scale for Adults (WAIS). The WAIS was further revised in 1981 and referred to as the Wechsler Intelligence Scale for Adults-Revised (WAIS-R, Wechsler, 1981). The WAIS-R was designed to measure intelligence in individuals ranging in age from 16 years 0 months to 74 years 11 months.

Following the same approach used to construct the adult scale, Wechsler derived the Wechsler Intelligence Scale for Children (WISC), published in 1949. A revised version, the Wechsler Intelligence Scale for Children-Revised (WISC-R) was developed in 1974. The WISC-R was further revised and published as the Wechsler Intelligence Scale for Children-III (WISC-III) (Wechsler, 1991). The WISC-III was designed to measure cognitive abilities in children ranging from 6 years 0 months to 16 years 11 months. Finally, a third intelligence scale, the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) was created in 1967. It was revised in 1989 and referred to as the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) (Wechsler, 1989). The WPPSI-R was developed for the assessment of children ranging in age from 3 years to 7 years 3 months.

The Wechsler scales serve a number of different purposes for clinicians and are often given a high degree of recognition in terms of the clinical information they yield. In fact, the Wechsler Scales are usually thought of as the standard for intellectual evaluation (Kaufman, 1990).

## Frequency of Use of Wechsler Scales

Surveys have indicated that the Wechsler intelligence scales are the most frequently used tests of intelligence. Harrison, Kaufman, Hickman, and Kaufman (1988) conducted a survey on test usage in adult assessments. Questionnaires were sent to members of four clinically oriented APA divisions and directors of APA accredited programs. Of 300 clinicians responding, the vast majority (97%) reported often using the WAIS or WAIS-R when a measurement of intelligence was required. All other intelligence tests were listed much less frequently. Overall, the WAIS-R was ranked first among the ten most commonly used tests in the 1980's. Seventy-four percent of clinicians responding to the survey believed the development of additional intelligence tests was unnecessary. The WAIS-R was perceived as yielding information of great importance and most respondents considered the norms and the theoretical soundness to be its primary strengths. Archer, Maruish, Imhof, and Piotrowski (1991) conducted a test usage survey sending questionnaires to practitioners working primarily with adolescents. The results of the survey revealed that the Wechsler scales of intelligence were the most commonly used tests, with 88% of survey respondents utilizing them in their practise. The Wechsler scales were also the most commonly cited test (91%) employed in standard clinical batteries.

Errors in the Administration of the Wechsler Scales

As the Wechsler scales are the most frequently used intelligence tests, most graduate schools provide specific instruction on these scales (Archer et al., 1991). However, studies have shown that a number of administration and scoring errors occur frequently in protocols given by psychology graduate students and professional psychologists (Bradley, Hanna, and Lucas, 1980; Levenson et al., 1988; Ryan, Prifiteria, & Powers, 1983; Slate & Jones, 1990a).

Many studies document specific concerns regarding administration inaccuracies on the Wechsler scales. The most common problems appear to include:

(i) numerous types of clerical, mathematical, and basal and/or ceiling errors,

(ii) errors in the evaluation, questioning and recording of responses on Verbal subtests - particularly Vocabulary, Similarities and Comprehension,

(iii) lack of adherence to the standardized directions found in the Wechsler manual for performance subtests - particularly Block Design and Picture Arrangement.

Clerical and Mathematical Errors. Numerous studies have shown that the Wechsler protocols of professional practitioners have a number of clerical and mathematical mistakes (Slate & Hunnicutt, 1988). Examples of clerical errors include (i) counting an item answered correctly after the discontinuance criterion has been fulfilled, (ii) illegible handwriting, and (iii) failure to assign points for assumed items that were not administered (Sherrets et al., 1979). Typical mathematical errors include incorrectly adding the raw and/or scale scores (Hajzler, 1987; Miller, et al., 1970; Ryan, et al., 1983) and inaccurate computation of the birth date (Sherrets, Gard, & Langer, 1979). Sherrets et al. (1979) examined the frequency of clerical errors on 200 WISC and WISC-R protocols. They found almost 89% of the examiners in the study made at least one mistake and 46.5% of the 200 protocols contained at least one error. The majority of clerical and mathematical errors included mistakes in obtaining the scale scores from the tables and incorrect addition of raw and scale scores.

Errors on Verbal subtests. Among errors frequently reported on the Wechsler scales are those related to the evaluation of responses on the Comprehension, Similarities, and Vocabulary subtests. Slate and Chick (1989), reported that the student administrators in their study (i) incorrectly allocated points on these subtests, (ii) failed to write responses down verbatim as stipulated in the manual, and (iii) questioned inappropriately by either failing to question as outlined in the manual or asking unwarranted questions. Slate, Jones, & Murray (1991) also found a number of errors on the Vocabulary, Comprehension and Similarities subtests including failure to record responses and the allocation of too many points to subject answers. Other studies have identified similar sources of error among the Verbal subtests (Fantuzzo, Sisemore, & Spradlin, 1983; Moon, Fantuzzo, & Gorsuch, 1986; Slate, & Jones, 1990c).

<u>Errors found in the Performance subtests.</u> Moon, et al., 1986 found Picture Completion and Block Design were the most inaccurately administered performance subtests. Errors included lack of adherence to the standardized directions for administration of the test and nonstandardized manipulation of the testing materials. Fantuzzo et al., (1983) employed an administration checklist, referred to as Criteria for Competent WISC-R Administration (CCWA), to evaluate WISC-R administrations by eight graduate students. They also reported frequent lapses in following standardized rules for presentation of the Block Design and Picture Arrangement subtests. Steward's (1987) results are similar to those reported by Fantuzzo et al. (1983). She used the WISC-

R Administration Observational Checklist (WAOC) to examine the degree and types of WISC-R administration errors performed by graduate students. Major sources of error included departures from the standardized directions and nonstandardized manipulations of Picture Arrangement and Block Design subtest materials.

Instruction on Administration of the Wechsler Scales

Wechsler test adminstration errors may arise as the result of examiner carelessness. At the same time, the inherent complexity of the administration procedures may contribute to variability in presentation. Stress, fatigue, boredom with testing, and time restraints due to excessive case loads have also been cited as possible sources of error in administration (Slate & Hunnicutt, 1988; Slate, Jones, Coulter, & Covert, in press). Alternatively, inadequate training procedures may diminish administration integrity. Typically, training involves a discussion of the administration and scoring procedures and a number of practise administrations. Generally, some of the practise administrations are evaluated and feedback given by an instructor or graduate assistant. Slate and Jones (1990b) found this procedure was unable to ensure competency and proficiency in administration. In their study, students given standard training averaged 11.3 mistakes on every WISC-R administration. None of the 217 protocols was without error. Furthermore, 5 practise administrations were not enough to increase accuracy and even on the 10th administration the number of errors was still high. Other studies examining traditional training procedures have also found them to be inadequate (Slate and Chick, 1989; Slate and Hunnicutt, 1988; Slate and Jones, 1990b). Slate, Jones, and Murray (1991) reported that practise administrations have at most a minimal influence on administration errors made by graduate students. Furthermore, Slate, Jones, Murray, & Coulter (under review) found that professional practitioners were even more likely to make mistakes on the WAIS-R compared to graduate students. Other studies have reported no significant differences in the number of mistakes made by students compared to professional practitioners (Sherrets et al., & Bradley et al., 1980). Inadequate training may yield assessors who do not possess adequate familiarization with administration details in the manual and/or who do not appreciate the importance of maintaining rigorous standardized procedures (Slate and Hunnicutt, 1988).

Impact of Administration and Scoring Errors

Wechsler (1981) emphasized that valid results depend upon strict adherence to standardized administration and scoring rules. Warren & Brown (1973) reported that 37% of the protocols given by a sample of graduate students contained errors that changed the Full Scale IQ as much as 5 points. Hajzler (1987) reported mistakes of 0 - 6 points on the Full Scale IQ with an average error of 1.56 IQ points. He also noted that a 6 point error when added to the standard error of the Full Scale IQ could potentially result in a mistake of up to 12 IQ points. Cummings and Moscato (1982) found that errors vary in terms of their impact on IQ scores. They noted that while administration and judgement errors were the most common source of error in their study, they produced the smallest changes in IQ scores. They argued that these errors tended to "cancel out" when the subtest score was tabulated. However, Cummings and Moscato found mathematical errors were able to significantly alter IQ scores. One examiner, for example, incorrectly added the Performance scale scores reporting a 54 instead of 44. The impact of administration and scoring errors should be evaluated further (Moon, Gorsuch, Blakely, and Fantuzzo, 1991). Such studies will help improve the validity of intelligence scales and should also be of benefit in designing better training programmes. In addition, findings may be used to calculate standard errors of measurement which incorporate administration and scoring errors.

Administration and scoring errors may ultimately result in an individual's exclusion or misplacement from a particular program or improper labelling of people who are not developmentally handicapped (Franklin, Stillman, Burpeau, & Sabers, 1982). Beasley, Lobasher, Henley, and Smith, (1988) expressed concern that errors of 5 points or more on the Full Scale IQ may influence a psychologist's decisions and conclusions with regards to a person's abilities and clinical progress. They noted that a falsely low IQ result for a child with phenylketonuria may be viewed as proof of a deterioration in intellectual performance. Conversely, a falsely high IQ result may hide an actual deterioration. Even a modest transformation in the IQ score can produce changes on an examinee's profile of cognitive strengths and weaknesses. Administration errors also raise legal and/or ethical issues surrounding professional competency and the Standards for Education and Psychological Tests as stipulated by the American Psychological Association (Moon et al., 1991).

Standard Administration of the WAIS-R Block Design

This study focused upon one administrative aspect of the Block Design, specifically, the presentation of the blocks as stipulated in the WAIS-R manual. The manual states that "the examiner should make sure that a variety of surfaces face up, that only one out of the four blocks has the red/white side facing up, and only three when nine blocks are used" (Wechsler, 1981, p.72). The author believes that many administrators randomly present the WAIS-R blocks instead of following the standardized procedure. Kaplan (1991) was of the same opinion that clinicians routinely toss the blocks for a random presentation of surfaces. The instructions to "scramble the blocks" found in the directions for the administration of a second trial for items 1 & 2 may be misleading, causing misunderstandings in the administration procedure. The standardized procedure for presentation may have been established in order to serve as a continual reminder to the examinees about the different sides on the blocks. Alternatively, it may serve to promote some consistency of block presentation without requiring examiners to present every block in a specific way.

This study compared standard administration of the blocks on the WAIS-R Block Design subtest with an extreme, non-standard presentation of the blocks. It served as a preliminary investigation of the impact of a typical error on WAIS-R administration. Furthermore, to determine whether intelligence level has a qualifying effect on non-standard administration of Block Design, the author attempted to obtain a sample with a wide range of IQs. It was predicted that:

(i) performance on Block Design would not differ for items presented in the standard (Wechsler's criteria) versus nonstandard (extreme) form,

(ii) there would be no differential effects of IQ on Block Design performance under extreme versus standard conditions of presentation.

## Method

### Subjects

Subjects were 60 adults, of which 35 were females and 25 were males. Forty-three of the subjects were volunteers from a first year university psychology pool. The other 17 subjects were obtained from two job readiness training programmes designed to teach work and occupational skills to adults who have not completed high school. The university subjects ranged in age from 18 years to 62 years and the subjects in the training programs ranged in age from 18 years to 57 years.

Procedure

The procedure for this study was approved by the Lakehead University Ethics Advisory Committee (see Appendix A).

WAIS-R Short Form

Each subject was given a four subtest short form of the Wechsler Adult Intelligence Scale-Revised employing the Similarities, Arithmetic, Picture Completion, and Digit Symbol subtests. This specific WAIS-R short form is referred to as Kaufman's 'quick tetrad' (Kaufman, 1990). The particular subtests were chosen by Kaufman because of their fast administration time and good prediction of Full Scale IQ. The tetrad has a reported mean reliability coefficient of .93 and a mean validity coefficient of .95. Therefore, it compares very favourably to Silverstein's V-A-BD-PA tetrad and Reynold's I-A-PC-BD tetrad with respect to psychometric properties (Kaufman, Ishikuma, and Kaufman-Packer, 1991).

WAIS-R Block Design

Each subject was given the Block Design subtest following

Kaufman's quick tetrad. The Block Design subtest was administered using a standardized and nonstandardized (extreme) presentation of the blocks. The WAIS-R manual stipulates that a variety of sides be shown and that the red/white side be facing up only on one of the blocks when four are used, and only on three blocks when all nine are given. In the extreme presentation, all blocks were provided with the full red side facing up. The extreme presentation was chosen for two related reasons: (1) to maximize the difference between standard and non-standard presentation of blocks, and (2) to avoid the variability in block face presentation that would arise by scrambling (ie., sometimes this procedure would comply with the standardized presentation by chance).[1] The form of presentation to each subject alternated with each Block Design item. If a subject failed the first trial on items 1 or 2, a second trial was given in accordance with the instructions in the WAIS-R manual. In this situation, a subject received the same form of block presentation as on their first trial of the item. The administration procedure was also counterbalanced by altering the presentation of the extreme or standardized arrangement for the first trial between subjects.

[1] This issue is taken up further in the discussion.

In total, 30 subjects received a standard presentation on the first item and alternate subsequent items. The other 30 subjects received the extreme presentation on the first item and every other item that followed.

## Results

The Age and IQ distributions for the total sample are presented in Tables 1 and 2. There was a good representation of subjects by age between 18 and 44 years in the study, but only 4 subjects with ages between 45 and 64. The mean age of the sample was 30.4 years. The IQ distribution was well represented within the low average to high average range. There were also three subjects in the borderline category and one in the superior range. The mean IQ of the sample using the short form test was 98.6 which closely approximates the mean IQ of the adult population.

Table 1

*Age Distribution of Subjects*

| Age of Subjects | Number of Subjects |
|---|---|
| 18-24 | 21 |
| 25-29 | 13 |
| 30-34 | 9 |
| 35-39 | 7 |
| 40-44 | 6 |
| 45-49 | 1 |
| 50-54 | 0 |
| 55-59 | 2 |
| 60-64 | 1 |

Table 2

*IQ Distribution of Subjects*

| IQ | Number of Subjects |
|---|---|
| 70-74 | 2 |
| 75-79 | 1 |
| 80-84 | 5 |
| 85-89 | 9 |
| 90-94 | 5 |
| 95-99 | 9 |
| 100-104 | 8 |
| 105-109 | 6 |
| 110-114 | 9 |
| 115-119 | 5 |
| 120-124 | 1 |

Effects of the Standardized versus Nonstandardized presentation on indices of Block Design performance

Table 3 summarizes the results of comparing the standard and nonstandard presentations on a variety of Block Design performance indices. Each index of performance and the statistical analysis are described in detail below.

Raw Scores:

The total Block Design raw scores for standardized item presentations and for nonstandardized (extreme) item presentations were tabulated separately for each subject. A dependent measures t-test indicated no significant difference between the mean raw score obtained on designs administered in the standardized manner ($M$ = 16.03, $SD$ = 5.98) and the mean raw score obtained for designs with an extreme presentation ($M$ = 16.87, $SD$ = 5.80), $t(59) = -1.26$.

Trials/Items Passed:

There were five subjects who failed either item 1 or 2 on their first attempt. Therefore, according to the manual, they were given a second trial or chance to complete the design. The form of presentation (ie. standard versus nonstandard) was not changed for the second trial. Thus, the number of trials (attempts) and the number of items (designs)

Table 3

Performance Indices for Standard versus Nonstandard Presentations

| | Standard | | Nonstandard | |
|---|---|---|---|---|
| | M | SD | M | SD |
| Raw Score | 16.03 | 5.98 | 16.87 | 5.80 |
| Trials/Items Passed | 3.60 | 1.08 | 3.83 | 0.96* |
| Trials Failed | 0.90 | 1.09 | 0.63 | 0.88* |
| Items Failed | 0.87 | 1.00 | 0.60 | 0.81** |
| Time | 25.15 | 11.61 | 27.54 | 10.53 |

* $p<.10$
** $p<.05$

are technically two different variables. However, when considering the 'passes', the number of trials passed by these five subjects is equivalent to the number of items passed. The mean number of passed standardized trials/items was 3.60 (SD = 1.08) and the mean number of passed nonstandardized trials/items was 3.83 (SD = 0.96). A dependent measures t-test indicated that this difference was not significant at the .05 level, $t(59) = 1.73$, $p<0.10$.

Trials/Items Failed:

When examining the difference between the mean number of failures for the two presentation formats, the number of failed trials and the number of failed items were considered separately. These measures are not the converse of trials/items passed since some subjects reached the ceiling level before all nine designs were given. When examining the failed trials each attempt by the subjects was counted. Therefore, subjects could receive two failures if they failed both trials on items 1 or 2. This occurred for only two subjects. The mean number of failed standardized trials (M = .90, SD = 1.09) did not differ significantly from the mean number of failed nonstandardized trials at the .05 level, (M = .63, SD = 0.88), $t(59) = 1.96$, $p<0.10$. When the number of items failed was examined, only the final attempts on items 1 and 2 were considered.

If a person passed the second trial, they received a pass for the item. Three subjects were in this situation. A dependent measures t-test revealed that the mean number of standardized items failed ($\underline{M}$ = 0.87, $\underline{SD}$ = 1.00) was significantly greater than the mean number of nonstandardized items failed, ($\underline{M}$ = 0.60, $\underline{SD}$ = 0.81), $\underline{t}(59) = 2.05$, $\underline{p} < 0.05$.

Time:

The mean of the times taken for the successful assembly of the designs in the standardized and nonstandardized presentations were calculated for each subject. A t-test revealed no significant difference between the mean of the subject mean time taken to successfully assemble standardized items ($\underline{M}$ = 25.15, $\underline{SD}$ = 11.61) and the mean of the subject mean time taken to successfully assemble the nonstandardized designs ($\underline{M}$ = 27.54, $\underline{SD}$ = 10.53), $\underline{t}(58) = -1.22$.

Frequency with which the last item passed was a standard or non-standard presentation:

The last item that each subject passed was examined to determine whether the item was presented in a standard or non-standard fashion. A Chi-Square analysis indicated that there was no significant difference between the number of subjects who passed a standardized presentation last versus the number of subjects passing a non-standardized presentation last, $\chi^2(1, \underline{N} = 60) = 2.4$

Reliability Measure

The split-half reliability coefficient was calculated for the Block Design items using the unequal-length Spearman-Brown technique. This coefficient was calculated for the 30 subjects who received standardized presentations on items 1, 3, 5, 7, and 9 versus nonstandardized presentations on items 2, 4, 6, and 8. Another coefficient was calculated for the 30 subjects who received nonstandardized presentations on items 1, 3, 5, 7, and 9 versus standardized presentations on items 2, 4, 6, and 8. These coefficients were respectively 0.89 and 0.81 and therefore were similar to the split-half coefficient of 0.87 reported in the WAIS-R manual.

Relationship between IQ and performance indices on Standardized versus Non-standardized presentations:

The analysis proceeded by examining the relationship between the short form estimate of Full- Scale IQ (SFIQ) and the difference between indices of performance for standard versus nonstandard presentation. It was important to investigate whether certain IQ levels (for example lower

IQs) were more/less affected by the different presentation formats and whether there was any consistency to such possible effects.

Raw Score and SFIQ:

The Block Design raw score for items with standardized presentations was subtracted from the Block Design raw score for items with nonstandardized presentations for every subject. A Pearson product-moment correlation indicated that the signed difference in the raw scores for standardized versus nonstandardized item presentations was not significantly correlated to the SFIQ, $r(58) = 0.15$.

Passed trials/items and SFIQ:

The number of standardized trials/items passed was subtracted from the number of nonstandardized trials/items passed to obtain a difference score for each subject. Again, it is worth noting that trials are equivalent to items when considering passes. A Pearson product-moment correlation revealed that the signed difference between the number of standardized and nonstandardized trials/items passed was not significantly related to the SFIQ, $r(58) = 0.10$.

Trials failed and SFIQ:

The number of nonstandardized trials failed was subtracted from the number of standardized trials failed to obtain a difference score for every subject. A Pearson product-moment correlation also revealed that the signed difference between the number of standardized and nonstandardized trials failed was not significantly related to the SFIQ, $\underline{r}(58) = 0.08$.

Items failed and SFIQ:

A difference score was calculated for each subject by subtracting the number of nonstandardized items failed from the number of standardized items failed. A Pearson product-moment correlation indicated that these signed difference scores were not significantly correlated to the SFIQ, $\underline{r}(58) = 0.12$.

Time and SFIQ:

The mean of each subject's times for successful assembly of nonstandardized items was subtracted from their mean time for successful assembly of standardized item presentations. The Pearson product-moment correlation coefficient was calculated between these signed time differences and the SFIQ. No significant relationship was found, $\underline{r}(57) = -0.18$.

Last item passed and SFIQ:

All subjects were assigned a code number of 1 or 2 depending upon whether the last item passed on the Block Design subtest was presented in the standard or nonstandard fashion respectively. A point biserial correlation coefficient was then calculated between this dichotomous variable and SFIQ. No significant relationship was found, $r(58) = 0.17$.

Relationship between Age and performance indices for standard versus nonstandard presentations:

There were no hypothesized relationships between age and performance indices on standard versus nonstandard presentation of the blocks. However, this variable was investigated by calculating correlation coefficients between age and each of the performance indices examined in the analysis with SFIQ. No significant relationships were found.

## Discussion

### Impact of an Extreme Presentation

The results of the study revealed that a non-standard extreme method of presenting the WAIS-R blocks had a negligible impact on Block Design performance. Key indices of performance such as raw score and time for successful assembly revealed no significant differences in performance when standard presentation was compared to nonstandard presentation. In addition, the analysis of the last item passed showed no significant difference in the frequency of standard versus nonstandard presentations. The only performance index to show a difference at traditional levels of significance was items failed. The mean number of standard presentation items failed was larger than the mean number of nonstandard presentation items failed. The result for trials failed was in the same direction but not strongly significant. These results need to be interpreted in the context of their magnitude and clinical significance. Significant differences were in the order of a fraction of an item or trial (.27 for items failed and .27 for trials failed). All indications are that these differences have negligible clinical impact as they do not translate into raw score differences. The same conclusions could be reached about trials/items passed.

The negligible effect of a nonstandard presentation was further bolstered by the analysis of performance indices with IQ and age. The differences between these indices for standard versus nonstandard presentations was not significantly related to SFIQ or the age of subjects. In addition, the split-half reliability coefficients were very close to the reported reliability coefficient for the Block Design subtest.

Implications for Random Presentation of Blocks:

In this study, all blocks were presented with the full red side facing up. This form was chosen for two related reasons. First, it was very extreme in appearance in an attempt to maximize the difference between standard and nonstandard presentations. It is believed that if the extreme presentation has negligible effects on performance, this serves as a good indication that a random presentation will be of the same magnitude or less.

The extreme presentation was also chosen in order to avoid the variability that would arise by scrambling the blocks (ie., sometimes this method would comply with the standardized presentation by chance). If four blocks, for example, are randomly scrambled, the probability of presenting the blocks in the standardized form is 0.30.[2] If nine blocks are used, as stipulated for designs 6-9, the probability of presenting the standardized form by chance is 0.26.[3] Random presentations of the blocks would not have resulted in a consistent experimental manipulation of the standardized procedure.

2. This probability was calculated by enumerating all the possible combinations of faces for four blocks.

3. The probability was calculated using two independent constraints; (i) three and only three blocks may be have the red/white side facing up, and (ii) of the six remaining blocks, not all may have the full white side or all have the full red side facing up.

Based upon the findings of this study, nonstandardized presentations of WAIS-R blocks, including random presentations, do not appear to seriously affect performance indices. However, this is a preliminary study on the evaluation of a departure from the standardized procedure for administration. The study does not account for all possible factors which may influence the results when blocks are presented randomly. For example, the standardization rule may help ensure that the blocks when presented, will not by chance closely resemble the target designs. Scrambling could make an item faster or easier to complete if the random configuration approximates elements of the target design. The influence of such an effect on performance would not be detected using the extreme presentation. A future study might try presenting approximations of the designs to see if this significantly affects test results. The influence of a nonstandard presentation on various populations should be considered as well. Would the results be the same if this study was conducted with children using the WISC-III or WPPSI-R? Various clinical populations, such as those with a neurological or memory impairment may also be studied. For these groups, the standard presentation which requires that a variety of surfaces face up, could serve as a continual reminder of the different sides on the blocks.

Conclusions:

The results of this study indicate that a nonstandard presentation of WAIS-R blocks has a negligible impact on performance. Clinicians who randomly present the blocks may take comfort in this fact. However, it would be prudent to adhere to the standard administration procedure since this will ensure uniform conditions. This study also can not account for all possible effects of a nonstandard presentation. Additional studies on this and other possible sources of error on the subtests of all Wechsler scales should be performed. Realizing the impact of errors on Block Design and other subtests will help develop a realistic sense of the current standard error of measurement connected with test scores. Hanna, Bradley, & Holen, (1981) suggested that the calculation of a composite error variance may be done by incorporating various sources of error including the common administration and scoring errors. Hanna et al. (1981) advocate that

> it seems far more prudent for practitioners to recognize the existence of these several sources of error and to make educated guesses concerning their magnitude than to ignore them. Sticking our heads in the sand will not make

them go away (p.375).

The problem with errors on the Wechsler scales continues to cause researchers to work on improving training and scoring procedures. Examiners are continuously cautioned to remain prudent. It may be wise to take additional steps to protect people from the impact of these errors. Computer programs included in the purchase of Wechsler scales which calculate the chronological age, tabulate scores, and convert raw scores into scale scores may eliminate some sources of error. Mandatory workshops could also be set up to keep administrators informed of error sources. As many researchers have noted, examiners must be kept aware of the impact of administration and scoring errors on their psychometric tools. The importance of this goal can be realized when considering tests such as the Wechsler scales which are proclaimed to contribute greatly to important clinical decisions.

References

Archer, R. P., Imhof, E. A., Maruish, M., & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. Professional Psychology: Research and Practice, 22 (3), 247-252.

Beasley, M. G., Lobrasher, M., Henley, S., & Smith, I. (1988). Errors in computation of WISC and WISC-R intelligence quotients from raw scores. Journal of Child Psychology and Psychiatry, 29, (1), 101-104.

Bradley, F. O., Hanna, G. S., & Lucas, B. A. (1980). The reliability of scoring the WISC-R. Journal of Consulting and Clinical Psychology, 48 (4), 530-531.

Cummings, J. A., & Moscato, E. M. (1982). WISC-R errors of state certified school psychologists. Unpublished manuscript, Institute for Child Study, Indiana University, Bloomington, IN.

Fantuzzo, J. W., Sisemore, T. A., & Spadlin, W. H. (1983). A competency-based model for teaching skills in the administration of intelligence tests. Professional Psychology; Research and Practice, 14, 224-231.

Franklin, M.R., Stillman, P. L., Burpeau, M. Y., & Sabers, D. L. (1982). Examiner error in intelligence testing: are you a source? Psychology in the Schools, 19, 563-569.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our head out of the sand. Journal of School Psychology, 19 (4), 370-375.

Harrison, P. L., Kaufman, A. L., Hickman, J. A., Kaufman, N. L. (1988). A survey of tests used for adult assessment. Journal of Psychoeducational Assessment, 6, 188-198.

Hajzler, D. J. (1987). Clerical errors in the scoring of the WISC-R. The Australian Educational and Developmental Psychologist, 4, 24-27.

Kaplan, E. (1991). The process approach to the clinical use of the WAIS-R. Paper presented at Ontario Psychological Association conference, Toronto, Ontario.

Kaufman, A.S. (1990). Assessing adolescent and adult intelligence. Boston: Allyn & Bacon, Inc.

Kaufman, A. S., Ishikuma, T., Kaufman-Packer, J. L. (1991). Amazingly short forms of the WAIS-R. Journal of Psychoeducational Assessment, 9, 4-15.

Levenson, R. L., Golden-Scaduto, C. J., Alosa-Karpas, C. J., & Ward A. W. (1988). Effects of examiner's education and sex on presence and type of clerical errors made on WISC-R protocols. Psychological Reports, 62, 659-664.

Miller, C., & Chansky, N. (1972). Psychologists' scoring of the WISC-R protocols. Psychology in the Schools, 9 (1), 144-152.

Miller, C. K., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. Psychology in the schools, 7, 190-193.

Moon, G. M., Gorsuch, R. L., & Fantuzzo, J. W. (1986). Teaching WAIS-R administration skills: Comparison of the MASTERY model to other existing clinical training modalities. Professional Psychology: Research and Practice, 17 (1), 31-35.

Moon, G. M., Gorsuch, R. L., & Fantuzzo, J. W., & Blakey, W. A. (1991). Frequent WAIS-R administration errors: An ignored source of inaccurate measurement. Professional Psychology: Research and Practice, 22 (1), 256-258.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R Journal of Consulting and Clinical Psychology, 51 (1), 149-150.

Sattler, J. S. (1990). Assessment of Children- (3rd Edition). San Diego: Jerome M. Sattler, Publisher.

Sherrets, S., Gard, G., & Langer, H. (1979). Freqency of Clerical Errors on WISC Protocols. Psychology in the Schools, 16 (4), 495-496.

Slate, J. R., & Chick, D. (1989). WISC-R Examiner Errors: Cause for Concern. Psychology in the Schools, 26 (1), 78-84.

Slate, J. R., & Hunnicutt, L. C. (1988). Examiner Errors on the Wechsler Scales. Journal of Psychoeducational Assessment, 6, 280-288.

Slate, J. R., & Jones, C. H. (1990a). Examiner Errors on the WAIS-R: A Source of Concern. The Journal of Psychology, 124 (3), 343-345.

Slate, J. R., & Jones, C. H. (1990b). Student error in administering the WISC-R: Identifying problem areas. Measurement and Evaluation in Counseling and Development, 23, 137-140.

Slate, J. R., & Jones, C. H. (1990c). Identifying Students' Errors in Administering the WAIS-R. Psychology in the Schools, 27, 83-87.

Slate, J. H., Jones, C. H., Coulter, C, & Covert, T. L. (in press) Practitioner's administration and scoring of the WISC-R: Evidence that we do err. Journal of School Psychology.

Slate, J. H., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the WAIS-R: An empirical evaluation of practice administrations. Professional Psychology: Research and Practice, 22 (5), 375-379.

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (under review). Practitioners err in administering and scoring the WAIS-R.

Stewart, K. J. (1987). Assessment of Technical Aspects of WISC-R Administration. Psychology in the Schools, 24 (3), 221-228.

Warren, S. A., & Brown, W. G. (1973). Examiner scoring errors on individual intelligence tests. Psychology in the Schools, 10, 118-121.

Wechsler, D. (1958). The measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1981). Wechsler Intelligence Scale for Adults-Revised. New York: The Psychological Corporation.

Wechsler, D. (1989). Wechsler Preschool and Primary Scale of Intelligence-Revised. New York: The Psychological Corporation.

Wechsler, D. (1991). Wechsler Intelligence Scale for Children-III. New York: The Psychological Corporation.