DECREASING SCORING ERRORS ON WECHSLER SCALE VOCABULARY, COMPREHENSION, AND SIMILARITIES SUBTESTS

Michele L. Linger

Certificate of Approval:

Steven G. LoBello, Ph. D.
Chairman
Professor
Psychology

Glen E. Ray, Ph. D.
Professor
Psychology

Peter Zachar, Ph. D.
Professor
Psychology

Roger A. Ritvo, Ph. D.
Vice-Chancellor for
Academic Affairs

DECREASING SCORING ERRORS ON WECHSLER SCALE VOCABULARY, COMPREHENSION, AND SIMILARITIES SUBTESTS

Michele L. Linger

A Thesis

Submitted to

The Graduate Faculty of

Auburn University Montgomery

In Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Montgomery, Alabama

May 9, 2005

DECREASING SCORING ERRORS ON WECHSLER SCALE VOCABULARY, COMPREHENSION, AND SIMILARITIES SUBTESTS

Michele L. Linger

Permission is granted to Auburn University Montgomery to make copies of this thesis at its discretion, upon the request of individuals or institutions and at their expense. The author reserves all publication rights.

Michele L. Linger
Signature of Author

May 9, 2005
Date

Copy sent to:

Name Date

VITA

Michele L. Linger, daughter of Donald and Carol Muth, was born March 4, 1974 in Kingston, New York. She attended Columbia-Greene Community College in Hudson, New York and graduated with an Associates Degree in Liberal Arts in May 1994. She attended Auburn University Montgomery in Montgomery, Alabama and graduated with a Bachelors Degree in Science in Psychology in May 2001. She entered Graduate School, Auburn University Montgomery, in June 2003.

THESIS ABSTRACT

DECREASING SCORING ERRORS ON WECHSLER SCALE VOCABULARY, COMPREHENSION, AND SIMILARITIES SUBTESTS

Michele L. Linger

Master of Science

(B.S., Auburn University Montgomery, 2003)

42 Typed Pages

Directed by Steven G. LoBello

Research has found that examiner error occurs during the administration and scoring of the Wechsler intelligence tests. It has been consistently reported that these errors occur more frequently on the Vocabulary, Comprehension, and Similarities subtests, which require more examiner judgment in administration and scoring. The purpose of this study was to reduce or eliminate examiner errors by focusing on how to correctly administer and score these subtests. Participants were enrolled in a graduate course in individual intelligence testing where they received concentrated instruction, practice scoring exercises, an objective test of the testing procedures, practice test administrations and scoring, and performance feedback. The dependent variables were the number of scoring errors made on the Vocabulary, Comprehension, and Similarities subtests. Scoring errors declined significantly for each of the subtests between the first and second sets of three intelligence tests administered by the student examiners, however individual examiner differences were a factor for both the Comprehension and the Similarities subtests.

AKNOWLEDGEMENTS

I would like to thank Dr. LoBello for his support and direction at each stage of this project. I would also like to thank Drs. Glen E. Ray and Peter Zachar for their insights and suggestions. Thanks also need to go to Haley Murray for her assistance in helping me to complete this project. A special thank you is given to my friends (Kurt and Shellie Bergo, Gina Williams, Melinda Krammer, and Emily DeBray), my parents, and especially my husband, Chris, for providing a shoulder to cry on, and for their never-ending support and encouragement. Lastly, I would like to thank my three daughters (Kathryn, Makayla, and Keily) for giving up some of their quality time so that I could continue in my academic aspirations.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

Decreasing Scoring Errors on Wechsler Scale Vocabulary, Comprehension, and Similarities Subtests

Because the Wechsler Scales are among the most frequently administered tests of intelligence, "it is imperative that they derive from competent administration and scoring" (Miller & Chansky, 1972, p. 152). In addition, "examiner errors affect Full-Scale IQ scores so as to create an adverse affect on both reliability and validity of test scores" (Slate, Jones, & Murray, 1991, p. 375). Slate and Jones (1990a) found that students learning to administer the WAIS-R made errors on 98% of their protocols and when these errors were corrected, the errors on 81% of the protocols resulted in changes in Full Scale IQs. Patterson and Slate (1995) found that "students administering the WAIS-R overestimated IQs on 71% of the protocols and underestimated IQs on 7% of the protocols, with 19% of the protocols having overestimates of four or more IQ points" (p.3). Belk, LoBello, Ray, and Zachar (2002) found examiner error to be associated with the incorrect assignment of IQ classifications on 11% of the WISC-III protocols used in the study.

Ryan, Prifitera, & Powers (1983) found that "regardless of one's experience level in psychological testing, scoring errors occur frequently and detract from the accuracy of the WAIS-R IQs" (p. 150). Franklin, Stillman, Young-Burpeau, & Sabers (1982) also using the WAIS-R found that even among certified professionals there is variation among scores. Bradley, Hanna, and Lucas (1980) found practitioners' calculated IQ values, on a sample of WISC-R protocols, varied by as much as 8 IQ points. Slate, Jones, Coulter, and Covert (1992) found that "practitioners committed errors on all 56 WISC-R protocols used in the study" (p.78). Klassen & Kishor (1996) reported that practitioners made

about 7 clerical errors regardless of which version of the WISC (WISC-R or WISC-III) was being administered. Depending upon where the errors are made and the extent of the errors, IQ scores could be miscalculated which in turn could affect IQ classifications for examinees.

Conner and Woodall (1983) found that the total number of errors made by student examiners, significantly decreases with experience in administration and scoring of the WISC-R. Platt, LoBello, Zachar, and Ray (2005) looked at the effects of providing students with feedback following a set of administered tests and found "practice along with feedback reduces errors, but does not eliminate them" (p. 16). Slate et. al.(1992) believed that "feedback had to occur immediately and must be specific in order to be effective, which is virtually impossible to do due to testing procedures" (p. 378). This is because the typical testing course requires that students turn in several protocols at once, not allowing for immediate feedback after each test is completed.

Irrespective of the edition or version of the Wechsler scales (adult or child), errors may be classified into three categories. Klassen and Kishor (1996) defined these as 1) administration error, which is deviating from the standardized instructions and test procedures; 2) scoring error, which involves incorrect point assignment to a test response; and 3) clerical error, which involves mistakes in addition, score conversions, and other clerical tasks associated with testing. Patterson and Slate (1995) believed that examiner errors were related to a misunderstanding of test instructions. This includes failing to query when clarification is needed in order to correctly assign points. Franklin et al. (1982) found examiner error to include failing to award the correct points to responses

and discontinuing subtests too early. They stated that in order to protect the validity of the Wechsler tests, examiner error must at be decreased.

The relationship between extended test administration and scoring practice has been difficult to assess in the literature. Studies indicate that experienced practitioners make substantial errors when administering the Wechsler Scales (Slate & Jones, 1990b; Slate et. al.,1992; Whitten, Slate, Shine, and Raggio, 1994). Klassen and Kishor (1996) found that even after 18 months of using the WISC-III, practitioners' error rates did not significantly change. In training student examiners, it has also been reported that student examiners do not improve from one test administration to the next (Slate, Jones, & Murray, 1991; Belk et. al., 2002). Patterson and Slate (1995) found that students still made errors and were not proficient even after several administrations of the WAIS-R. Slate, Jones, & Murray (1991) investigated the value of giving a large number of tests, as well as the effect learning to administer the WISC-R before the WAIS-R. They found that increasing the number of test administrations was not related to a significant decrease in examiner errors. They also found that learning to administer the WISC-R prior to administering the WAIS-R was actually associated with an increase in errors because of differences in the score conversion procedures that existed between those two scales.

Platt et. al. (2005) demonstrated that small but significant improvements in examiner proficiency are found when errors are measured after practice and feedback, rather than after each individual test administration. This is consistent with the findings of Connor and Woodall (1983) and raises the possibility that previous findings of the ineffectiveness of practice may be the result of inappropriate methodology. Specifically,

expecting error reductions after each test, and without corrective feedback, is probably unrealistic.

Methods of Reducing Examiner Errors

Competency-based training was the focus of the Moon, Fantuzzo, and Gorsuch (1986) study. This involved studying the WAIS-R manual, giving a test, receiving feedback on the administration, attending a lecture based upon the dangers of administration errors, and the administration of a second test. The examiners' ability to follow standardization criteria outlined by Wechsler, on the second administration, was assessed using the Criteria for Competency WAIS-R Administration (CCWA). The CCWA had a total of 17 subdivisions, which included Vocabulary, Comprehension, and Similarities subtests. They found improvement across all subdivisions, which included the Vocabulary, Comprehension, and Similarities subtests.

Blakey, Fantuzzo, Gorsuch, and Moon (1987) extended the Blakey et. al. (1985) study by investigating the use of peer mediation in competency-based training. In their study, "one student administered the first five items of each subtest to another student, while a third student evaluated the student examiner according to the Criteria for Competent WAIS-R Administration manual" (p.18). They were also instructed to study the Criteria for Competency WAIS-R Scoring manual. They were then required to take a test covering scoring criteria, which allowed them to become peer trainers. "In this study, it was found that peer-mediated competency-based training procedures could be used effectively to train students to administer and score the WAIS-R competently" (p. 18). It was also that found that these students improved the accuracy of scoring the Comprehension and Similarities subtests.

Slate and Jones (1989) found that increasing the amount of instruction time, focusing directly on how to administer and score the WISC-R, describing in detail the most commonly made errors, and providing a detailed list of rules for avoiding these errors, results in fewer examiner errors and a decrease in the quantity of Full Scale IQs needing to be corrected.

Thompson and Hodgins (1994) developed the Compu-Check Form to check clerical and computational procedures. Clerical errors are due to incorrect written information on the protocol or incorrect item scoring and computational errors involve errors in adding numbers and chronological age. Thirty-four percent of test protocols contained scoring errors prior to the use of the Compu-Check Form. Use of the form was associated with improvement in clerical and computational accuracy, with 10% of the protocols containing scoring errors after the Compu-Check Form was implemented. However the Compu-Check Form was used only to eliminate computational errors.

Subtests Vulnerable to Examiner Error

The Vocabulary, Comprehension, and Similarities subtests, unlike other Wechsler scale subtests, require examiners to make judgments based on scoring criteria and assign 0, 1, or 2 points to each response. Miller and Chansky (1972) found that the greatest amount of scoring variance among raters occurs on the Vocabulary, Comprehension, and Similarities subtests. Examiners tended to disagree about scoring when the responses to test items were unclear, and particularly when the response to be scored was not included in the scoring examples provided in the test manual.

Slate et al. (1991) evaluated student administered WAIS-Rs and reported that the Vocabulary, Comprehension, and Similarities subtests were the most error-prone because

examiners misunderstand the scoring criteria and do not recognize the wide variation in responses or response quality. Slate et. al. (1992) also reported that examiner errors were caused by assigning too few or too many points to responses, or were caused by inappropriate follow-up questioning.

The Vocabulary, Comprehension, and Similarities subtests have consistently been the most error-prone subtests on the Wechsler Scales for children and adults. Miller and Chansky (1972) found that examiners differed in scoring from 8-12 items for each of these subtests when scoring a single WISC-R protocol. In other studies of students and experienced examiners using a variety of Wechsler Scales, Vocabulary, Comprehension, and Similarities subtests usually are most prone to scoring error (Slate & Jones, 1990b; Slate et. al., 1991; Slate et. al., 1992; Belk et. al., 2002). Some subtests are either less vulnerable to error, or the primary errors occur during test administration and are not discernable from protocol review. For example, very small error frequencies were found on the WISC-III Object Assembly, Digit Span, and Arithmetic subtests (Belk et. al., 2002). It makes little sense to concentrate educational or research efforts on the remediation of errors that rarely occur and have minimal impact on test scores.

This study was designed to determine if examiner error could be reduced or eliminated on the WAIS-III and the WISC-IV, Vocabulary, Comprehension, and Similarities subtests by focusing on how to correctly administer and score these subtests. To determine whether or not there was a reduction or elimination of examiner error, individual differences among examiners were included in the model as covariates and removed from the error term. This was done to control for initial differences among students' testing abilities. The hypothesis is that concentrated instruction, practice

scoring exercises, and an objective test of the testing procedures, along with practice test administrations and performance feedback will significantly reduce scoring errors on the Vocabulary, Comprehension, and Similarities subtests on the Wechsler Adult Intelligence Scale, Third Edition and The Wechsler Intelligence Scale for Children, Fourth Edition.

Method

Participants

Twelve master's-level psychology students enrolled in the intelligence testing class at Auburn University Montgomery participated in the study. Participants gave consent to participate (see appendix A). The consent form signed by examinee test participants (see appendix B) included a section indicating that their test results may be used for research purposes. Of the twelve participants, eleven were females. The male participant and one of the female participants were African American. The remaining ten participants were White. All participants had previously taken a course in general psychometric theory, but none of them had previous experience in the administration of individual intelligence tests. Due to a limited number of test kits, half of the students learned to administer the WAIS-III first and half first learned to administer the WISC-IV.

Design and Procedure

This research project is a one-group, pretest-posttest repeated measures design. Students received typical instruction on how to administer and score both the WAIS-III and the WISC-IV. Typical instruction included an overview of the scales and topics such as calculating chronological age and the various I.Q. values, following directions in the manual, subtest administration and scoring, as well as instructions on how to conduct an evaluation and write psychological reports. Several activities related to the administration and scoring of the Vocabulary, Comprehension, and Similarities subtests were designed to bring students to the same level of proficiency. Students completed two in-class scoring exercises for either the WISC IV or the WAIS-III. The scoring exercise that was given was dependent upon which test the student would first be administering.

The scoring exercises were designed to provide practice in scoring of the Vocabulary, Comprehension, and Similarities subtests before scoring actual protocols. The scoring exercises were completed and then discussed during a single class period. Discussion of the scoring exercises focused on explaining the scoring for questions by providing the basic criteria for how to score the questions.

The students were required to pass a ten question multiple-choice test covering the technical criteria for scoring the Vocabulary, Comprehension, and Similarities subtests. The students were told in advance that they would have to take and pass a multiple-choice test covering the scoring criteria for Vocabulary, Comprehension, and Similarities as outlined in the manual. This was done to ensure that all students would read and review the manual prior to beginning their actual tests. The students got the majority of the questions correct with scores ranging from seven to nine.

The students then administered and scored three test protocols. Students received written feedback (see appendices C & D) about any mistakes made on the above test with specific emphasis on Vocabulary, Comprehension, and Similarities subtests. Students also received feedback on at least one of the first three protocols they had administered and scored. Feedback was immediately provided to only one protocol because it was believed that the protocol was representative of the remaining two protocols. Once these activities were completed and feedback had been provided, the students administered and scored a second set of three test protocols.

A reliability check procedure was used to ensure that reliability existed among the raters. The reliability check procedure was fashioned after the procedure that was used in Platt et. al. (2005) and Belk et. al (2002). Two test protocols for both the WAIS-III and

the WISC-IV were randomly selected. Two graduate students independently reviewed and corrected the scoring of each item on the Vocabulary, Comprehension, and Similarities subtests. The two raters then compared errors and raw scores for the Vocabulary, Comprehension, and Similarities subtests. Differences in scores were reconciled and all test protocols in the sample were rescored to reflect the reconciliation. Two more test protocols for both the WAIS-III or the WISC-IV were randomly selected. The raters proceeded in the above fashion until there were no differences in scores on two sets of two consecutive protocols for both the WAIS-III and the WISC-IV. Once this occurred, the reliability check procedure was discontinued. A single rater determined errors on the remaining protocols, but all had been checked and corrected during the reliability procedure. Mean error rates were calculated for the Vocabulary, Comprehension, and Similarities subtest for each group of protocols (determined by the order of administration).

Setting and Materials

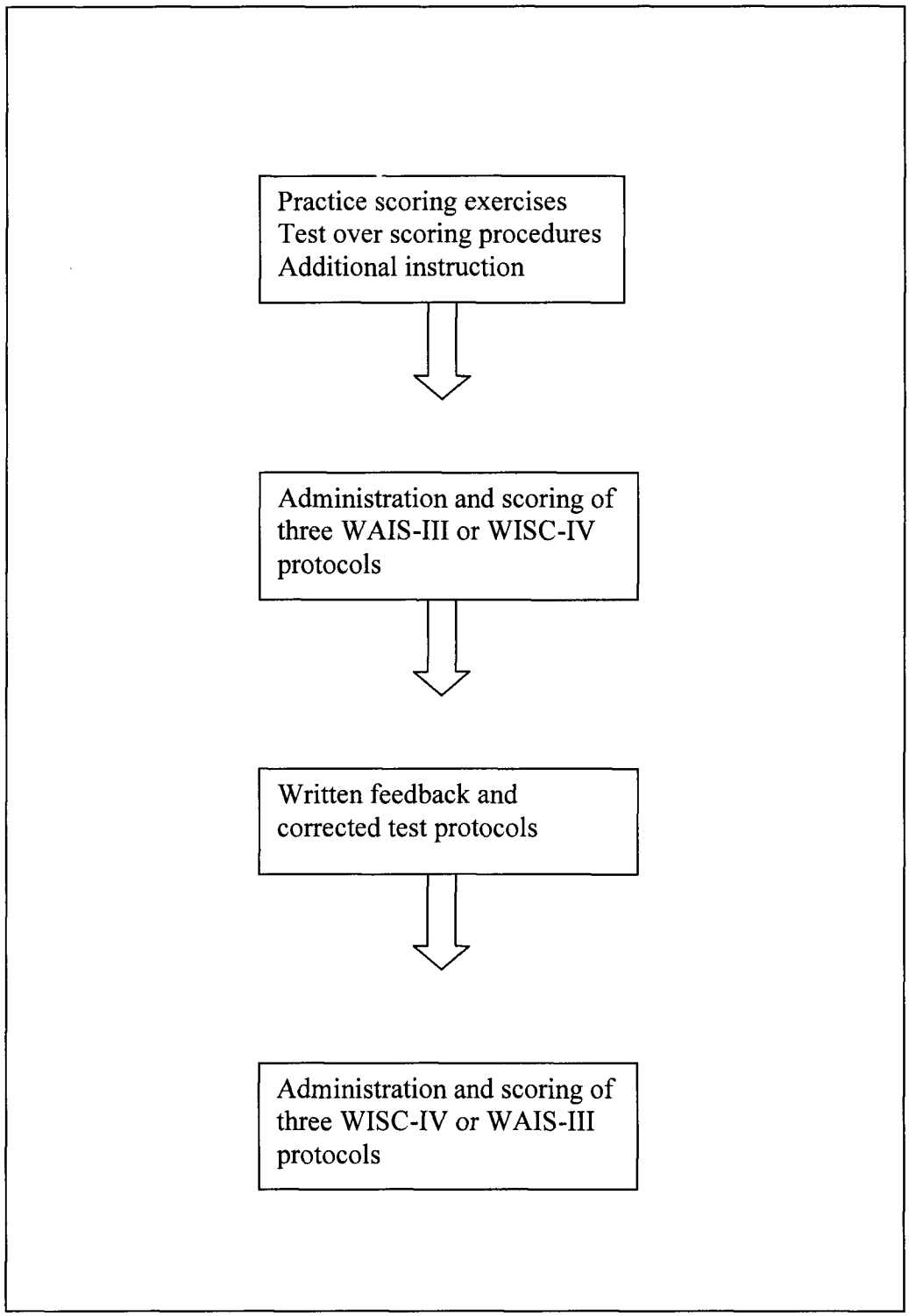
The study took place at Auburn University Montgomery in the classroom where the Intelligence Testing class was taught. Standard WAIS-III and WISC-IV test kits, protocols, and manuals were used. Checklists for WAIS-III and WISC-IV Vocabulary, Comprehension, and Similarities subtests were used to determine the number and types of errors made on test protocols. The WAIS-III Checklist was used in a previous research study that focused on the effects of practice and feedback with the WAIS-III and the WISC-III (Platt et. al. 2005). The WISC-IV Checklist was made for this study, but was modeled after the WISC-III Checklist that was used in the Platt (2005) study on the

effects of practice and feedback (checklists are not included in the appendix in order to protect the confidentiality of test content).

Data Analysis

The study design was a one-group pretest-posttest repeated measures design with the number of scoring errors on Vocabulary, Comprehension, and Similarities as the dependent variable. Data were analyzed with repeated measures analysis of variance. Variation due to individual student differences was counted as between subjects variance and removed from t error term. The study hypothesis predicted a decrease in scoring errors on the three subtests between pre-test and post-test for each individual subtest and for combined errors on the three subtests.

Figure 1. Study flow chart.



Results

This study was designed to investigate the effectiveness of strategies to decrease the number of scoring errors made by examiners on the Vocabulary, Comprehension, and Similarities subtests for the WAIS-III and the WISC-IV. A 2(test sets) x 12(individual participants) repeated measures design was used where the dependent variable was the number of scoring errors made on each subtest. The hypothesis for this study was that, as a result of practice and feedback, scoring errors would decline significantly from the first to the second set of tests administered. The SAS general linear model program was used to analyze the data. Variation due to individual differences among examiners was included in the model as a covariate and removed from the error term.

The reliability check described in the previous section was used to ensure that the determination of errors was in general agreement with the judgments of a second rater. The two raters reached the agreement criteria of no differences in subtest raw scores or scaled scores after jointly reviewing eight sets of two protocols for each test.

Means and Standard Deviations

The means and standard deviations for the Vocabulary, Comprehension, and Similarities subtests for the first and second set of tests administered were calculated. Table 1 presents the means and standard deviations of the number of errors made on the Similarities, Vocabulary, and Comprehension subtests on the first and second sets of tests administered.

Table 1. Mean Errors (with Standard Deviations) for Vocabulary, Comprehension, and Similarities for Test Sets 1 and 2

Subtest	Test Set 1	Test Set 2
Vocabulary	4.91 (3.54)	3.61 (2.86)
Comprehension	5.88 (4.96	3.83 (3.23)
Similarities	4.19 (3.26)	2.86 (3.59)

Vocabulary

Data were analyzed using separate models for each subtest with number of errors on a given subtest as the dependent variable. A repeated measure ANOVA was used in each case. The first model determined if the number of errors that occurred on the Vocabulary subtest from the first set of tests administered to the second set declined significantly, while controlling for individual differences among examiners. The repeated measures ANOVA revealed that the mean number of errors decreased significantly from the first to the second set of tests. The partial η^2 associated with Vocabulary subtest errors was .16, indicating that the reduction of errors on the Vocabulary subtest accounted for 16% of the variance in the model, excluding variance attributable to the between subjects variation. Table 2 summarizes the results of the ANOVA, and also indicates that individual examiner (between subjects) differences were not a significant source of variation in the model, although this component of the model accounted for 35.7% of the total variance.

Table 2. Repeated Measures Analysis of Variance Source Table for Vocabulary Errors

Source	SS	df	MS	F	p
Test Set	30.68	1	30.68	4.54	.04
Examiners	107.5	11	9.77	4.51	.22
Error	162.33	24	6.76		
Total	300.51	36			

Comprehension

The next analysis determined if there was a significant decline in the number of errors that occurred on the Comprehension subtest from the first set of tests administered to the second set while controlling for individual differences among examiners. A significant decrease in the number of errors from the first set of tests to the second set of tests was revealed by the ANOVA procedure, which is summarized in Table 3. The partial η^2 associated with the reduction of errors in the Comprehension subtest was .28. Unlike the analysis of the Vocabulary subtest, the results indicate that a significant source of variation in this model was due to individual examiner differences. The η^2 associated with examiner (between subjects) differences was .64, or 64% of the total variance of the model.

Table 3. Repeated Measures Analysis of Variance Source Table for Comprehension Errors

Source	SS	df	MS	F	p
Test Set	76.06	1	76.06	9.52	.003
Examiners	473.3	11	43.03	5.39	.0003
Error	191.67	24	7.98		
Total	741.03	36			

Similarities

The third model determined if, while controlling for individual differences, there was a significant decline in the number of errors that occurred on the Similarities subtest from the first set of tests administered to the second set. The repeated measures ANOVA, summarized in Table 4, revealed that there was a significant decrease in the number of errors from the first set of tests to the second set of tests The partial η^2

associated with the reduction of errors in the Similarities subtest was .21. Individual examiner differences were a significant source of variation in the model, accounting for 54.6% of the total variance.

Table 4. Repeated Measures Analysis of Variance Source Table for Similarities Errors

Source	SS	df	MS	F	P
Test Set	32.00	1	32.0	6.33	.02
Examiners	184.67	11	16.8	3.32	.0007
Error	121.33	24	5.1		
Total	338	36			

Scaled Scores

The next step was to determine if the Scaled Scores for the Comprehension and Vocabulary subtests changed after correcting errors to the raw scores. Using the same repeated measures analyses as with the number of errors, it was found that there were no significant differences between the original Scaled Scores and the Scaled Scores based on corrected raw scores. The values in Table 5 are the mean Scaled Scores of Vocabulary, Comprehension, and Similarities subtests before and after the correction of errors. Inspection of the means in Table 5 indicates that error corrections on these three subtests had negligible impact on the subtest scaled scores.

Table 5. Mean Subtest Scaled Scores (Standard Deviations) for Vocabulary, Comprehension, and Similarities Subtests Before and After Raw Score Correction.

SUBTEST	BEFORE CORRECTION	AFTER CORRECTION
Vocabulary	11.40 (2.48)	11.39 (2.45)
Comprehension	11.48(3.19)	11.28 (2.91)
Similarities	11.42 (2.78)	11.42 (2.87)

Discussion

This study focused on decreasing the number of scoring errors made by student examiners on the Vocabulary, Comprehension, and Similarities subtests on the WAIS-III and the WISC-IV. It was believed that the number of errors would decrease by requiring students to take a practice protocol, providing additional focus on instructions, and requiring them to pass a test regarding the criteria for the scoring procedures for the Vocabulary, Comprehension, and Similarities subtests.

Twelve master's level psychology students participated in the study. The initial course activities were designed to introduce, describe, and provide experience in scoring the Vocabulary, Comprehension, and Similarities subtests. Further, the purpose of these activities was to bring all students to the same level of competence before administering Wechsler scales to volunteer participants. Students were provided with lectures, in-class practice scoring exercises, discussion of the criteria for scoring items on the scoring exercises, and a ten question multiple-choice test.

Results of this study indicate that the mean number of errors on the Vocabulary, Similarities, and Comprehension subtests declined after practice and feedback, which is supportive of the study hypothesis. Earlier research findings that suggested no improvement with practice suffered methodological weaknesses that concealed the effects of practice and feedback (Platt, et al, 2005). Specifically, some studies did not use repeated measures designs, and others examined for changes after each the protocol, rather than immediately after practice and feedback.

Individual differences among examiners were a significant factor in the models assessing decreases in errors on the Similarities and Comprehension subtests, but not the

Vocabulary subtest. These findings suggest that the procedures used to bring all students to the same level of competence before they begin testing are effective for the Vocabulary subtest, but not for the others. The implications of this finding are that additional or different procedures are needed to increase initial scoring competence on the Similarities and Comprehension subtests. Other possibilities are that trait or trait-like examiner differences (e.g., Conscientiousness) may be involved in tendency to make errors. However, the findings of this study indicate that it is possible to train students to a standard of like competence on the Vocabulary subtest. More effective and innovative teaching methods should be explored before attributing problems to difficult-to-change internal characteristics.

Scaled scores based on uncorrected raw scores were no different from scaled scores based on corrected raw scores. The examiner errors, when corrected, could either add points to the raw score or take them away. The net impact of the errors on many protocols would be negligible once the raw scores were converted to scaled scores. It is encouraging to know that these errors cannot generally affect Full Scale IQ scores. Past research (Belk, et. al, 2002 & LoBello & Holley, 1999) has shown that other errors, such as miscalculating chronological age or using optional subtests in calculating IQ values may cause very large changes in IQ. Despite the minimal impact of errors on scaled scores, it is still important to train examiners to eliminate scoring errors on the Vocabulary, Comprehension, and Similarities subtests. Errors on these subtests may be an indicator of poor understanding of testing procedures, which could extend to other areas of test administration and scoring.

This study used a small sample of convenience, limiting the generalizability of the results. However, the unit of measure was the 36 test protocols for both the WISC-IV and the WAIS-III, which provided sufficient data for the study. Another problem that this study faced was the inability to differentiate which intervention (practice or feedback) caused the errors to decline. When small classes are used in research, and where the primary purpose is to teach and provide supervised practice, compromises in experimental procedures will usually be required. Participants are self selected to be in the class, and the small number of participants limits the ability to operate multi-group experiments. Treatment dismantling studies would be quite difficult in the classroom context. Randomly assigning students to a control group provided less than optimal instruction and experience would be unethical.

The loss of experimental control is offset somewhat by the ecological validity of conducting research in a 'real world' setting. Graduate students in intelligence testing classes are usually trained in small classes, and usually are motivated to perform well, if not because of genuine interest, then because of a desire to earn a passing grade. Studies of this type could be performed with volunteer participants who would agree to receive instruction on how to score intelligence test items, and then practice these activities until mastery was attained. This procedure would offer the opportunity to increase sample sizes, utilize random assignment, and determine the relative effectiveness of various training strategies. However, volunteer participants would likely not be graduate students in psychology, and would probably have limited interest in and motivation to learn and practice the test scoring procedures. A study requiring participants to attend multiple training and practice sessions would also likely suffer attrition over time. No students in

the current study dropped the intelligence-testing course until after completing the testing described in this study.

The direction of future research is clear. Effective procedures that bring all students to a similar level of competence in scoring the Similarities and Comprehension subtests need to be developed. The decreases in errors found in this study, while statistically significant, were still too numerous after practice and feedback. In future studies, students should be given more frequent feedback, perhaps after each test protocol is administered, and should not initiate additional testing until they have received recommendations for improving practice.

This study is the first to address examiner errors on subtests of the Wechsler Intelligence Scale for Children – Fourth Edition. Additionally, the findings from this study are consistent with research findings, such as the Platt et. al. (2005) study, which found that practice administration and feedback was effective in reducing overall examiner errors on WAIS-III and WISC-III. Platt et. al. (2005) and the present study contradict earlier research findings, which reported that practice and feedback did not lead to the development of competence among student examiners learning to administer individual intelligence tests. This study provides a glimpse of what may be done to decrease the number of errors made by student examiners on selected Wechsler scale subtests. The goal of this research program is to find the most effective way to teach graduate students to competently score the Wechsler scales Vocabulary, Similarities, and comprehension subtests, and to reduce errors as much as possible. Although this study indicates that the methods employed were most effective with Vocabulary subtest errors,

future studies are needed to discover similarly effective means of reducing errors on the Comprehension and Similarities subtests.

References

Belk, M. S., LoBello, S. G., Ray, G. E., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment,* 20, 292-304.

Blakey, W. A., Fantuzzo, J. W., Gorsuch, R. L., & Moon, G. W. (1987) A peer-mediated, competency-based training package for administering and scoring the WAIS-R. *Professional Psychology: Research and Practice,* 18(1), 17-20.

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.

Conner, R. & Woodall, F. E. (1983). The effects of experience and structured feedback on WISC-R error rates made by student-examiners. *Psychology in the Schools,* 20, 376-379.

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

Klassen, R. M. & Kishor, N. (1996). A comparative analysis of practicioners' errors on WISC- R and WISC-III. *Canadian Journal of School Psychology,* 12(1), 35-43.

Lobello, S. G. & Holley, G. (1991). WPPSI-R administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment,* 17, 15-23.

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

Moon, G. W., Fantuzzo, J.W., & Gorsuch, R. L. (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.

Patterson, M. & Slate, J. R. (1995). The effects of practice administrations in learning to administer and score the WAIS-R: a partial replication. *Educational & Psychological Measurement,* 55(1), 32-37.

Platt, T. L., LoBello, S. G., Zachar, P., & Ray, G. E. (2005). *The effects of practice and organized feedback on scoring and administration errors on Third Edition Wechsler scales.* Unpublished manuscript, Auburn University Montgomery, Montgomery, Alabama, Dept Psychology.

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

Slate, J. R. & Jones, C. H. (1989). Can Teaching of the WISC-R Be Improved? Quasi-Experimental Exploration. *Professional Psychology: Research and Practice,* 20(6), 408-410.

Slate, J. R. & Jones, C. H. (1990a). Identifying students' errors in administering the WAIS-R. *Psychology in the Schools,* 27, 83-87.

Slate, J.R. & Jones, C.H. (1990b). Examiner errors on the WAIS-R: A source of concern. *The Journal of Psychology,* 343-345.

Slate, J. R., Jones, C. H., Coulter, C., Covert, T. L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. *Journal of Social Psychology,* 30, 77-82.

Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale-Revised: an empirical evaluation of practice administrations. *Professional Psychology: Research and Practice,* 22(5), 375-379.

Thompson, A. & Hodgins, C. (1994). Evaluation of a checking procedure for reducing clerical and computational errors on the WAIS-R. *Canadian Journal of Behavioral Science,* 26(4), 492-504.

Appendix A

Informed Consent

Auburn University Montgomery
Psychology Department

You are invited to participate in a study of administrator errors on the individually administered intelligence tests. The purpose is to help determine the best way to teach individual intelligence testing skills to graduate students. You were selected for participation because you are currently enrolled in an Intelligence Testing class where you will be required to receive instruction and to administer individual intelligence tests.

As a member of the Individual Intelligence Testing class, you will be exposed to the teaching methods and exercises used in the study. Your consent for participation in the study covers only the inclusion of your test protocol data into the study sample. If you do not provide consent, your data will not be included in the study.

Any personally identifying information obtained during this study is confidential and will not be disclosed to third parties. A database will be constructed that will not include any identifying information other that basic demographic data (age, gender, etc.). Only numerical summaries and statistical analyses will be reported in research reports. Nothing will be reported that may lead to your personal identification.

Your decision whether to participate will not prejudice your future relations with Dr. LoBello, the Psychology Department, or Auburn University Montgomery. If you decide to participate, you are free to withdraw your consent and to discontinue participation at any time without penalty. Further, deciding to not participate, or to withdraw participation, will not affect your grade in the Individual Intelligence Testing class. If you decide to later withdraw from the study, any information collected will be excluded from analysis.

If you have any questions, you may ask either Dr. LoBello or Michele Linger. If you have any additional questions later, we will be happy to answer them. You may call Michele Linger at 262-2850 or Dr. LoBello at 244-3309. You will be given a copy of this form to keep.

YOU ARE MAKING A DECISION WHETHER TO PARTICIPATE. YOUR SIGNATURE INDICATES THAT YOU HAVE DECIDED TO PARTICIPATE, HAVING READ THE INFORMATION PROVIDED ABOVE.

_____ _____

Date Time Respondent's Signature

_____ _____

Witness Print Respondent's name

_____

Investigator's signature

Appendix B

AUBURN UNIVERSITY OF MONTGOMERY
School of Sciences
Department of Psychology

Statement of Informed Consent

You are invited to participate in an exercise that will be of assistance in the training of a graduate student who is learning to administer individual intelligence tests. The purpose of this exercise is to give students experience in the administration, scoring, and interpretation of individual intelligence tests.

The administration of these tests can take anywhere from 1 ½ to 2 ½ hours and the entire procedure can usually be completeted in one session. The procedures pose no risks to your health or safety. However, you have the right to terminate your participation at any time. As a result of your participation, you will have a better understanding of the work of psychologists in clinical settings (if you are an introductory psychology student, your instructor may awarded extra credit for your participation).

Your responses to the test questions and tasks will remain confidential, as will all test forms. The student may submit a written report of your test results to the instructor, but your identity will not be contained in the report.

The individual who tests you is a graduate student in training. Because this is a training experience, scores obtained from the administration of the test will not be disclosed to you or any other third party. The policy of not disclosing test results allows students to obtain supervised experience under conditions that minimize any potential adverse consequences for you.

The course instructor is Dr. Steven LoBello, AUM, 210F Goodwyn Hall, 244-3309. You are encouraged to contact him if you have any questions or concerns about your participation in this exercise.

YOU MUST BE AT LEAST 18 YEARS OF AGE TO SIGN THIS FORM. IF UNDER AGE 18, PARENT OR GUARDIAN MUST SIGN IN ORDER FOR CONSENT TO BE VALID.

_____

Signature of Volunteer or Person Authorized to Sign for Volunteer — Date

_____

Witness

Appendix C

Checklist for **WISC-IV** Protocols

Name of Student_____ Birth Date of Client_____ Date of Test_____

1. Chronological age: CORRECT INCORRECT

ENTER NUMBER IN BLANKS BELOW

2. _____ subtest raw scores correct, but copied incorrectly to Score Conversion page of protocol.

3. Scoring errors on individual items caused _____ subtest scales scores to change value

4. _____ subtest scaled scores copied incorrectly from tables to front of protocol

5. _____ sum of subtest scaled scores copied incorrectly from Score Conversion Page to Profile Page.

6. Addition errors: Similarities YES NO
Vocabulary YES NO
Comprehension YES NO

7. Wrong norms tables used YES NO

****

8. Similarities score assigned by student _____
Recalculated Similarities score _____
Difference (+/-) _____

Vocabulary score assigned by student _____
Recalculated Vocabulary score _____
Difference (+/-) _____

Comprehension score assigned by student _____
Recalculated Comprehension score _____
Difference (+/-) _____

Appendix D

Checklist for **WAIS-III** Protocols

Name of Student_____ Birth Date of Client_____ Date of Test_____

1. Chronological age: **CORRECT INCORRECT ENTER NUMBER IN BLANKS BELOW**

2. _____ subtest raw scores correct, but copied incorrectly to Score Conversion page of protocol.

3. Scoring errors on individual items caused _____ subtest scales scores to change value

4. _____ subtest scaled scores copied incorrectly from tables to front of protocol

5. _____ sum of subtest scaled scores copied incorrectly from Score Conversion Page to Profile Page.

6. Addition errors: Similarities YES NO
Vocabulary YES NO
Comprehension YES NO

7. Wrong norms tables used YES NO

****

8. Similarities score assigned by student _____
Recalculated Similarities score _____
Difference (+/-) _____

Vocabulary score assigned by student _____
Recalculated Vocabulary score _____
Difference (+/-) _____

Comprehension score assigned by student _____
Recalculated Comprehension score _____
Difference (+/-) _____