Psychology in the Schools, Vol. 46(7), 2009
Published online in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/pits.20405

GRADUATE STUDENTS' ADMINISTRATION AND SCORING ERRORS ON THE WOODCOCK-JOHNSON III TESTS OF COGNITIVE ABILITIES

ERICA RAMOS AND VINCENT C. ALFONSO

Fordham University, Graduate School of Education

SUSAN M. SCHERMERHORN

Irvington Union Free School District

The interpretation of cognitive test scores often leads to decisions concerning the diagnosis, educational placement, and types of interventions used for children. Therefore, it is important that practitioners administer and score cognitive tests without error. This study assesses the frequency and types of examiner errors that occur during the administration and scoring of the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG). Data from 36 graduate students across 108 test records revealed a total of 500 errors across all records. Further analyses indicated three frequently occurring errors, including the use of incorrect ceilings, failure to record errors, and failure to encircle the correct row for the total number correct. The results of this study may be used to inform training programs so that appropriate steps can be taken to decrease the number of examiner errors on the WJ III COG and similar cognitive test batteries. © 2009 Wiley Periodicals, Inc.

The interpretation of cognitive test scores often leads to decisions regarding the diagnosis, educational placement, and types of interventions used for children, adolescents, and adults (Alfonso & Pratt, 1997; Braden & Alfonso, 2003). Therefore, it is important that the administration and scoring of cognitive tests are performed without error. However, many studies indicate that a variety of examiner errors occur during the administration and scoring of widely used cognitive tests that may compromise the scores (Alfonso & Pratt, 1997). Although many of these errors may be as simple as incorrect addition of raw score points or using the incorrect starting points according to the age of the individual, they can also be as critical as calculating and reporting inaccurate overall scores, such as the Full Scale IQ (FSIQ) of the Wechsler Scales (Alfonso, Johnson, Patinella, & Rader, 1998). Consequently, it seems critical that examiners be cognizant of such errors and the frequency with which they occur so that they may be reduced and the accuracy of the administration and scoring of these tests may improve. By decreasing the number of examiner errors in cognitive tests, practitioners will be able to make more valid diagnostic and intervention decisions for their clients.

Summary of Previous Research

Previous research that has investigated examiner errors on the various Wechsler Scales indicates that regardless of the examiners' level of training (professional or graduate student), a significant number of errors have been found across test records (Alfonso et al., 1998; Loe, Kadlubek, & Marks, 2007; Sherrets, Gard, & Langner, 1979; Slate & Chick, 1989; Slate, Jones, Coulter, & Covert, 1992). The most frequent types of errors revealed in studies on the Wechsler Scales include failure to record responses, reporting incorrect FSIQ, adding subtest scores incorrectly, incorrect point assignment, use of incorrect basal and ceilings, incorrect calculation of age, and a number of other less frequent errors (Alfonso et al., 1998; Alfonso & Pratt, 1997; Loe et al., 2007; Slate & Jones, 1990). Table 1 provides a summary of several studies that have assessed examiner errors on the various Wechsler Scales. The results of these studies demonstrate that regardless of the examiners' levels of experience (i.e., graduate students in training or practicing professionals), a high number

Correspondence to: Vincent C. Alfonso, Fordham University, Graduate School of Education, 113 W. 60th Street, New York, NY 10023. E-mail: alfonso@fordham.edu

Table 1
Summary of Results of Previous Studies Investigating Examiner Errors on the Various Wechsler Scales

Study	Sample	Instrument Investigated	Major Findings
Sherrets, Gard, & Langner (1979)	39 psychologists, interns, practicum students, school psychologists, and psychometricians	WISC	89% of examiners made at least one error; most common errors were in addition of scores
Slate & Chick (1989)	14 graduate students	WISC-R	All subtests were found to have some error; errors on 66% of the protocols resulted in changes to FSIQ
Slate & Jones (1990)	26 graduate students	WISC-R	An average of 11.3 errors per protocol; frequent errors included failure to record examinee responses, incorrect point assignment, and inappropriate questioning
Slate, Jones, Coulter, & Covert (1992)	9 certified psychological examiners	WISC-R	An average of 38.4 errors per protocol, including failure to record responses; errors on 81% of the protocols resulted in changes to FSIQ
Alfonso, Johnson Patinella, & Rader (1998)	15 graduate students	WISC-III	An average of 7.8 errors per protocol; frequent errors included failure to query, failure to record responses verbatim, reporting incorrect FSIQs, reporting incorrect VIQs, and incorrect addition of scores
Loe, Kadlubek, & Marks (2007)	17 graduate students	WISC-IV	An average of 25.8 errors per protocol; common errors were failure to query, assigning too many points to a response, failure to record an examinee's response, and inaccurate test composite scores, resulting in incorrect FSIQ and Verbal Comprehension Index

Note. WISC = Wechsler Intelligence Scale for Children; WISC-R = Wechsler Intelligence Scale for Children−Revised; FSIQ = Full Scale IQ; WISC-III = Wechsler Intelligence Scale for Children, Third Edition; VIQ = Verbal IQ; WISC-IV = Wechsler Intelligence Scale for Children, Fourth Edition.

of examiner errors occur during the administration and scoring of these cognitive tests, which often lead to the miscalculation of FSIQ scores.

Recommendations for Reducing Examiner Error

Results of previous studies have led to many recommendations in training programs to reduce the number of examiner errors on cognitive tests. Several researchers (e.g., Alfonso & Pratt, 1997; Oakland & Zimmerman, 1986; Piotrowski & Zalewski, 1993) have recommended increasing the length of the training course to 1 year and increasing the time commitment of instructors and students taking the course in cognitive assessment. For instance, Alfonso, Oakland, LaRocca, and Spanakos (2000) revealed that many school psychology training programs offered only one semester of instruction in cognitive assessment. In addition, with the emphasis on response to intervention, some programs may reduce semester hours in cognitive assessment to the bare minimum (Mather

& Kaufman, 2006a, 2006b). Alfonso and Pratt (1997) recommended that the number of practice administrations required of students should approximate five to six administrations per test if possible, and that performance objectives and checklists be available for the students as a method for providing feedback on their administration and scoring skills. Furthermore, they provided recommendations that can be implemented in the classroom setting that may enhance student performance on test administration, such as allowing opportunities for peer training and practice administrations, providing instruction on common errors in administration and scoring, and incorporating assessment skills into competency exams.

Woodcock-Johnson III Tests of Cognitive Abilities

Although previous studies have assessed examiner errors on the various Wechsler Scales, and to a lesser extent the Kaufman Assessment Battery for Children (e.g., Hunnicutt, Slate, & Gamble, 1990), no published studies have investigated examiner errors on the Woodcock-Johnson (WJ) series of cognitive tests. The most recent in the WJ series of cognitive tests is the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG; Woodcock, McGrew, & Mather, 2001). The WJ III COG is a revision of the Woodcock-Johnson Tests of Cognitive Abilities—Revised (WJ-R; Woodcock & Johnson, 1989) and measures specific and general cognitive abilities based on the Cattell-Horn-Carroll (CHC) theory of cognitive abilities (McGrew & Flanagan, 1998). The WJ III COG is an individually administered test of cognitive abilities that provides scores that represent most of the broad cognitive abilities subsumed by CHC theory. The WJ III COG uses an easel format and does not contain manipulatives, which helps streamline the administration procedures. The WJ III COG scoring procedures are simplified by the accompanying scoring program, which only requires the entering of raw scores for each subtest, and some additional data entry that is required for the accurate generation of derived scores. Although the scoring program allows for fast and accurate results of individual performance for this measure, it is not without limitations. For example, to obtain accurate derived scores, the examiner must first double-check any errors he or she may have made during the administration of the subtests. Furthermore, accurate calculation of raw scores by the examiner is a prerequisite to obtaining accurate derived scores from the scoring program.

The use of this measure in training programs and by a number of professionals has rapidly increased since its publication due to the streamlined administration and user-friendly electronic scoring system of the test (Braden & Alfonso, 2003). Considering the increasing use of the WJ III COG series in practice and training, approximately 210 graduate programs to date in various disciplines of psychology (Woodcock-Munoz Foundation, personal communication, September 19, 2007), it is important to determine the administration and scoring accuracy of examiners who use this measure. Identifying the types of examiner errors that occur most frequently will provide valuable information that may be used to prepare graduate students and practitioners alike for the administration and scoring of this instrument.

In 2002, Schermerhorn and Alfonso began collecting data regarding examiner errors of graduate students in the administration and scoring of 14 WJ III COG subtests. To collect data in a standardized fashion, they developed a checklist[1] to evaluate the types and frequency of errors of 22 graduate students across 78 test records. The checklist was adapted from a similar checklist used for the Wechsler Scales that describes common administration errors found in each subtest (Alfonso et al., 1998). This checklist and results of preliminary data collected by Schermerhorn and Alfonso were reported in Braden and Alfonso (2003), and indicated a number of examiner errors in the administration and

[1] The WJ III COG Administration and Scoring Checklist is available on request from the second author and is also found in Braden and Alfonso (2003, pp. 386–387).

scoring of the WJ III COG. More specifically, incorrect starting points were used on about 10% of the test records. Incorrect calculation of raw scores and subsequent entry of that incorrect raw score into the scoring program were found on approximately 14% of the test records in the Spatial Relations subtest. Moreover, circling of incorrect raw scores occurred on 7% of the test records in the Numbers Reversed subtest. Data analysis at that time indicated very few errors found on the descriptive information, observation checklist, questions on the cover sheet, administration of correct designated sample items, subtests, and correct versions of the subtest on each of the 78 test records. Because additional data have been collected since the Braden and Alfonso (2003) publication, this study expands on the work and results reported previously to assess further the types and frequency of examiner errors that occur during the administration and scoring of the WJ III COG.

Method

Participants

Participants were 36 graduate students (31 female, 5 male) attending an American Psychological Association–accredited and National Association of School Psychologists–approved School Psychology Program from a large metropolitan university. Thirty participants were White, 4 were Latino, 1 was Black, and 1 was Asian. The participants were graduate students enrolled in one of three cognitive assessment courses taught by the same instructor. Ten students were enrolled in the course in the first semester that data were collected, 12 during the second semester, and 18 during the third semester. Although 40 students in total were enrolled in the course over three semesters, only 36 students participated in the study; 4 students did not participate because they did not complete their assignments (i.e., test records) in a timely fashion. All 36 participants included in this study were first-year students who had not been previously exposed to any cognitive test, including the WJ III COG.

Procedure

The course instructor required the graduate students to administer and score the WJ III COG four times throughout one semester, resulting in a total of 144 test administrations across the 36 students. The instructor implemented recommendations found in the literature throughout the course. For example, as recommended in Slate and Jones (1990), after an in-depth lecture and class discussion on the background and development of the WJ III COG, as well as a description of the various subtests and test stimuli, students conducted one administration of subtests 1 to 7 and 11 to 17 of the WJ III COG to a fellow classmate and three other administrations to volunteers.[2] The practice WJ III COG administration with a classmate provided an opportunity for students to present feedback on administration skills and help reinforce those skills prior to administering the assessment to other volunteers. Moreover, the instructor provided the students with extensive verbal and written feedback about their performance on each administration before conducting their subsequent administration.

This descriptive study calculates the frequency and types of examiner errors most commonly committed on the WJ III COG. These data would be used to inform training programs and practicing professionals who use this measure so that fewer errors would be committed in practice. Therefore, the checklist first developed by Schermerhorn and Alfonso (2002) for use in course instruction and consisting of possible errors for each subtest on the WJ III COG was used in this study to determine the number and types of errors in each test record (e.g., calculating the raw score incorrectly). The

[2] These volunteers included children, adolescents, and adults because the WJ III COG spans ages 2 to 90+ years.

first test record for each student was not included in the data set because it served as a practice administration. Therefore, only three test records were included in the data set for each of the 36 students, resulting in a total of 108 WJ III COG test records.

Each of the 108 test records was then examined and rated independently by one of four raters. The raters were advanced-level graduate students who had previously taken the same course with the same instructor and had been trained by the instructor on how to use the checklist. All four raters used the standardized checklist procedures for each test record. Interrater reliability was not calculated because subjectivity was minimal with the use of the checklist. Raters were simply asked to report if the error had occurred, or if it was not applicable to the subtest.

Results

A total of 500 errors across 108 test records were recorded with a mean of 4.63 errors per test record made by the graduate students in training. It is important to note that 167 of the total errors (33%) were made on 5 test records, whereas 50 test records (46%) had either 0 errors or 1 error. Results indicate that the subtests in which the most errors occurred were Verbal Comprehension, Visual-Auditory Learning, General Information, and Retrieval Fluency (Table 2).

A review of the errors made by the students on each WJ III COG subtest indicated that the three most frequently occurring errors were the use of incorrect ceilings, failure to record examinee errors, and failure to encircle the correct row for the total number correct (Table 3). Examiners failed to adhere to the ceiling rules of the subtests and either tested beyond the ceiling or did not reach the required ceiling. Examiners also failed to record the incorrect responses of the examinees. Table 3 shows the percent of test records that contained incorrect ceilings in the Numbers Reversed, General Information, Auditory Attention, and Memory for Words subtests. The percent of test records for

Table 2
Frequency of Graduate Student Errors for Each Subtest Across 108 WJ III COG Test Records

Subtest	Number of Errors
Verbal Comprehension	59
Visual-Auditory Learning	52
Spatial Relations	38
Sound Blending	25
Concept Formation	35
Visual Matching	13
Numbers Reversed	40
General Information	50
Retrieval Fluency	45
Picture Recognition	41
Auditory Attention	31
Analysis Synthesis	27
Decision Speed	06
Memory for Words	32
Total	494

Note. There were six additional errors that occurred during the administration of the protocols that are not reflected in this table because the errors were not specific to any subtests. For instance, these errors included failure to complete identifying information and failure to answer additional questions.

Table 3
Percent of the Three Most Common Errors Across 108 WJ III COG Test Records

	Type of Error		
Subtest	Incorrect Ceiling	Failure to Record Examinee Errors	Failure to Encircle Correct Row
Verbal Comprehension	—	12	7.4
Visual-Auditory Learning	—	—	7.4
Spatial Relations	—	—	9.3
Sound Blending	—	10.2	—
Concept Formation	—	12	7.4
Visual Matching	—	—	6
Numbers Reversed	15.7	5.6	5
General Information	14.8	13	4
Retrieval Fluency	—	—	8.3
Picture Recognition	—	—	8.3
Auditory Attention	4.6	14	4.6
Analysis-Synthesis	—	7.4	4.6
Decision Speed	—	—	3
Memory for Words	5	10.2	4.6

Note. (—) indicates that the error was not applicable to the subtest or that it occurred less than 1% of the time.

which examiners failed to record errors on the Verbal Comprehension, Sound Blending, Concept Formation, Numbers Reversed, General Information, Auditory Attention, Analysis-Synthesis, and Memory for Words subtests is also found in Table 3, along with the percent of test records that examiners failed to encircle the correct row for the total number correct on all subtests. In addition, incorrect entry of subtests' raw scores into the software program occurred for a total of 51 times across the 108 test records. Each administration error can lead to incorrect raw and/or derived scores on the WJ III COG for the individuals who completed the measure.

The difference in the number of errors across the three test records for each graduate student was also assessed. The total number of errors for each of the 108 test records was assessed by summing up the errors in each of the three test records across all participants. A total of 191 errors occurred during the administration and scoring of the first test record, 176 errors occurred during the administration and scoring of the second test record, and 133 errors occurred during the administration and scoring of the third test record (see Table 4 for means and standard deviations). A within-subject repeated measures analysis of variance (ANOVA) was conducted to determine whether there was a significant difference in the number of errors committed across the three test administrations. Results indicated a significant difference between the first and third administrations ($F = 6.707$, $p < .05$). That is, significantly fewer errors were committed on the third test administration as compared to the first test administration.

Table 4
Means and Standard Deviations of Errors Across the Three Test Administrations

Administration	M	SD
First	5.31	7.30
Second	4.89	9.28
Third	3.69	6.45

Note. $N = 36$.

Discussion

Based on the data and results reported previously, it is evident that errors occur when graduate students in training administer and score the WJ III COG. For example, the use of incorrect ceilings, failure to record examinee responses verbatim, failure to circle the correct row for total number correct, and incorrect entering of raw scores into the software program may compromise individuals' scores, as well as subsequent diagnostic, placement, or intervention decisions.

When compared with similar studies on the Wechsler Scales, for example, it seems that the most frequently occurring errors on the WJ III COG are similar. For example, on the Wechsler Scales, examiners failed to record examinee responses verbatim and used incorrect basal and ceilings (Alfonso et al., 1998; Alfonso & Pratt, 1997; Slate & Jones, 1990). Another similarity between the results of this study and previous studies of the Wechsler Scales is the average number of errors across test records. However, because scoring of the WJ III COG is almost solely dependent on the scoring software program, it is critical that examiners enter the raw scores correctly and double-check their work before using the scoring program. Once the raw scores have been entered into the scoring program, errors may be more difficult to uncover than those on the Wechsler Scales because those calculations cannot be verified manually. In addition, errors seem to occur most frequently on the Verbal Comprehension, Visual-Auditory Learning, General Information, and Retrieval Fluency subtests. Therefore, examiners may benefit from more focus and practice on the correct administration and scoring of these subtests. Although many errors were found, our results also indicated 46% of the test records had either 0 errors or 1 error, which speaks to the ease of the administration procedures of the WJ III COG.

Two main limitations of this study were selection procedures of the sample and the lack of an experimental control group. The sample used in this study consisted of students from one university, who were trained by one instructor, which limits the generalizability of the results. Nevertheless, results of this study are more similar to results of previous studies than they are dissimilar. In addition, the number of participants is consistent with previous studies. Prior studies suggest that providing verbal and written feedback to trainees, increasing the number of practice administrations, and using competency-based training models may decrease the number of administration and scoring errors on intelligence tests (Alfonso & Pratt, 1997). However, the lack of an experimental control group prohibits our ability to determine whether any given method of instruction or practice mentioned previously could have reduced the number of errors found in the test records. Furthermore, it should be noted that the checklist used for data collection in this study did not allow raters to fill in any raw scores from the protocols, which in turn made it difficult to determine which types of errors had a significant impact on raw scores.

Given the aforementioned limitations of this study, suggestions for future research include the following:

1. Investigate the impact of incorrect raw scores on derived standard scores.
2. Examine administration and scoring errors of similar, commonly used measures, such as the Stanford-Binet Intelligence Scales, Fifth Edition (Roid, 2003), the Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004), and the Differential Ability Scales, Second Edition (Elliot, 2007).
3. Investigate instruction of graduate training courses in cognitive assessment to determine commonly used instructional methods in training programs.
4. Include an experimental group in the study in order to determine the most effective means of instruction in cognitive assessment courses.

Future research may also consider including practicing professionals in the sample to determine whether there are any differences in the types of errors made by practicing professionals versus graduate trainees, and how knowledge of these errors can benefit practicing professionals.

Limitations, notwithstanding, the results of this study may be useful for instructors of cognitive assessment courses. It is recommended that instructors consider these findings when teaching the WJ series of tests and take steps, such as double-checking the protocols and subtest scores, practicing a number of administrations, and providing feedback on these administrations, to try to reduce the number of errors. Perhaps with closer examination of test records by students and instructor, the frequency of these common errors can be decreased. These results can be shared with students so that common pitfalls of administration and scoring of the WJ III COG can be avoided. Furthermore, the WJ III COG Administration and Scoring Checklist may be used as a teaching tool and an instructor aid, and can be adapted for other specific uses to assess common errors. These results also shed some light on the need for more careful consideration of the administration and scoring of tests such as the WJ III COG on behalf of the examiner when conducting assessments. It is imperative that examiners understand the many implications of test scores and do what they can to make this and all other aspects of assessment procedures as accurate as possible.

References

Alfonso, V. C., Johnson, A., Patinella, L., & Rader, D. E. (1998). Common WISC-III examiner errors: Evidence from graduate students in training. Psychology in the Schools, 35, 119–125.

Alfonso, V. C., Oakland, T., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. School Psychology Review, 29, 52–64.

Alfonso, V. C., & Pratt, S. I. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 326–344). New York: Guilford.

Braden, J. P., & Alfonso, V. C. (2003). The Woodcock-Johnson III Tests of Cognitive Abilities in cognitive assessment courses. In F. A. Schrank & D. P. Flanagan (Eds.), WJ III clinical use and interpretation: Scientists-practitioner perspectives (pp. 377–415). San Diego: Academic Press.

Elliot, C. D. (2007). Differential Ability Scales, Second Edition (DAS-II). San Antonio, TX: Pearson Education.

Hunnicutt, L. C., Slate, J. R., & Gamble, C. (1990). Examiner errors on the Kaufman Assessment Battery for Children: A preliminary investigation. Journal of School Psychology, 28, 271–278.

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Assessment Battery for Children, Second Edition (KABC-II). San Antonio, TX: Pearson Education.

Loe, S. A., Kadlubek, R. M., & Marks, W. J. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. Journal of Psychoeducational Assessment, 25, 237–247.

Mather, N., & Kaufman, N. (Eds.). (2006a). Special issue, part one: Integration of cognitive assessment and response to intervention [Special issue]. Psychology in the Schools, 43(7).

Mather, N., & Kaufman, N. (Eds.). (2006b). Special issue, part two: Integration of cognitive assessment and response to intervention [Special issue]. Psychology in the Schools, 43(8).

McGrew, K., & Flanagan, D. (1998). The Intelligence Test Desk Reference: Gf-Gc cross-battery assessment. Needham Heights, MA: Allyn & Bacon.

Oakland, T. D., & Zimerman, S. A. (1986). The course on individual mental assessment: A national survey of course instructors. Professional School Psychology, 1, 51–59.

Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved Psy.D. and Ph.D. clinical psychology programs. Journal of Personality Assessment, 61, 394–405.

Roid, G. H. (2003). Stanford-Binet Intelligence Scales, Fifth Edition (SB5). Itasca, IL: Riverside.

Schermerhorn, S. M., & Alfonso, V. C. (2002). Graduate student administration and scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. Unpublished manuscript.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. Psychology in the Schools, 16, 495–496.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. Psychology in the Schools, 26, 78–83.

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

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 School Psychology, 30, 77–82.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III Tests of Cognitive Abilities. Itasca, IL: Riverside.