Pergamon

Journal of School Psychology, Vol. 40, No. 2, pp. 167–176, 2002

Printed in the USA
0022-4405/02 $–see front matter

PII S0022-4405(02)00091-2

Rater Agreement on IQ and Achievement Tests Effect on Evaluations of Learning Disabilities

Robert G. Van Noord and Frances F. Prevatt
Department of Human Services, 215 Stone Building, Florida State University Tallahassee, Tallahassee, FL 32306, USA

Protocols from 110 evaluations utilizing the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) and the Woodcock/Johnson Tests of Achievement-Revised (W/J-R) were scored by two different raters to determine (a) whether subtests with more difficult levels of scoring yield lower interrater correlation coefficients, (b) whether scoring errors on subtests affect broad score estimates, (c) the effect of expertise of rater on scoring errors, and (d) whether scoring errors affect a learning disability determination based on IQ/achievement discrepancy. Scoring errors were found on almost 25% of Comprehension and Vocabulary subtests; however, the effect of these scoring errors was minimal. About 42% of Writing Samples subtests had scoring errors, resulting in a mean change of 1.75 points on the Broad Written Language Cluster subtest. On the WISC-III, but not the W/J-R, there were significantly more errors made by inexperienced testers. Scoring errors resulted in two cases in which learning disability determination would be changed. Overall, the study corroborates previous findings of strong interrater reliability on most subtests of common IQ and achievement tests and indicates that novice scorers are not likely to make scoring mistakes that will significantly impact an IQ/achievement discrepancy-based documentation of learning disability. © 2002 Society for the Study of School Psychology. Published by Elsevier Science Ltd

Keywords: WISC-III, W/J-R, Interrater reliability, Assessment, Learning disabilities.

Public Law 94-142 (Education of all Handicapped Children Act, 1975), Public Law 101-476 (Individuals with Disabilities Act, 1990), the IDEA 97 revisions (Individuals with Disability Education Act Amendments, 1997), and the Rehabilitation Act of 1973 mandate that students meeting established criteria for specific learning disorders receive reasonable accommodations in order to ensure the provision of appropriate public education curricula. More than one-half of all children receiving special education/ESE services

Received 17 July 2001; received in revised form 24 October 2001; accepted 9 November 2001.

Address correspondence and reprint requests to Frances F. Prevatt, Department of Human Services, 215 Stone, Florida State University Tallahassee, FL 32312, USA. Phone: (850)-644-9445; fax: (850)-644-4335; E-mail: fprevatt@coe.fsu.edu

nationwide are diagnosed with a learning disorder (Shinn, Good, & Parker, 1999). Specific guidelines for the identification of learning disabled students vary from state to state; however, most criteria are consistent with the broad guidelines established by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition-Text Revision (DSM-IV-TR, American Psychiatric Association, 2000). The DSM-IV-TR (2000) provides criteria for Specific Learning Disorders in the following three areas: reading, mathematics, and written expression. In addition to an evaluation of academic achievement and sensory deficits, the DSM states that learning disorders are diagnosed when the individual's "achievement on individually administered, standardized tests in reading, mathematics, or written expression is substantially below that expected for age, schooling, and level of intelligence ... Substantially below is usually defined as a discrepancy of more than 2 standard deviations between achievement and IQ" (p. 49).

The IDEA revision of 1997 states that "the team finds that a child has a severe discrepancy between achievement and intellectual ability ..." (300.541). Using a discrepancy criteria has been widely debated, with difficulties noted based on regression to the mean (making it easier to achieve a discrepancy if you have a higher IQ score), lack of interchangeability of tests, and questionable use of full scale IQ scores (Dumont, Willis, & McBride, 2001). However, because many states utilize an IQ/achievement discrepancy criterion as part of the overall decision making, scoring accuracy is crucial to LD determination.

INTERRATER RELIABILITY

The issue of interrater reliability, the need for multiple examiners to score the same quantitative test items in a consistent manner, represents a fundamental concern to test reliability and validity. High interrater agreement is essential to control for error in test administration.

The Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) and the Woodcock/Johnson Tests of Achievement-Revised (W/J-R; Woodcock & Johnson, 1989) represent two of the instruments most frequently used by school psychologists to determine the presence of a learning disability (Stinnett, Havey, & Oehler-Stinnett, 1994). Because scoring of achievement and intelligence tests can be quite laborious and subject to examiner error, the present investigation deals with interrater reliability, the effect of scoring errors on broad score estimates, the relationship between scorer expertise and scoring errors, and the effect of scoring errors on the determination of specific learning disabilities.

Studies of the interrater reliability of the WISC-III have found lower interrater reliability coefficients for specific subtests, most notably Similarities, Vocabulary, and Comprehension (Cuenot & Darbes, 1982; Shrout & Fleiss,1979). While interrater discrepancy is contingent to

some degree on test subject variability, several studies have indicated that test variance is associated with examiner errors (Alfonso, Johnson, Patinella, & Rader, 1998; Park, 1999). These studies have primarily targeted populations of graduate students at various stages of test administration training. Alfonso et al. (1998) found that a high prevalence of scoring errors persisted throughout several administrations of the WISC-III by the students in training. Their findings indicate that many of these errors consisted of lack of familiarity with established scoring criteria, such as failure to query subject responses and failure to transcribe subjects' responses in a verbatim manner. Research conducted by Park (1999), also involving graduate student trainees, corroborated these findings, and also cited frequent errors in subtest timing and reading instructions verbatim. Park's study identified significantly elevated instances of errors resulting from student trainees' specific administration of the WISC-III Comprehension, Block Design, and Picture Arrangement subtests.

The W/J-R battery has received some criticism of its psychometric stability. Although a new version has been released (W/J-III), the Writing Samples, Letter–Word Identification, Passage Comprehension, Word Attack, and Applied Problems remain the same. With regard to scoring errors, Simpson and Halpin (1995) found some deviation from established standardization procedures on the W/J-R (i.e., reducing the number of item errors required to obtain a ceiling, or expanding the range of items within which several consecutive incorrect responses must occur) to be statistically negligible to test outcomes. Interrater reliability of the Writing Samples test (unchanged from the WJ-R to the WJ-III), based on the standardization sample, is reported as a "typical interrater correlation of about 0.90" in one study and a "typical intercorrelation of about 0.98" in a second study (McGrew & Woodcock, 2001, pp. 43–44).

The present study evaluated the effects of rater reliability of common IQ and achievement tests on subsequent learning disorder eligibility determinations, particularly with respect to difficulty level of individual subtests and expertise of the scorer. This study had the following objectives: (1) To determine whether IQ and achievement subtests with more difficult levels of scoring yielded significantly lower interrater correlation coefficients than subtests generally thought to be easier to score. Subtests hypothesized to result in lower interrater correlation coefficients included Similarities, Vocabulary, Comprehension, and Writing Samples. (2) To determine whether scoring errors on subtests affected broad score estimates, such as cluster scores and Full Scale IQ. (3) To determine whether level of expertise was associated with scoring errors, and, if so, whether specified subtests were more difficult to score. (4) To determine whether scoring errors affected the number of students meeting the criteria for a specific learning disability.

METHOD

Participants

Examinees. This study evaluated test protocols from 110 subjects who were individually administered the WISC-III and W/J-R at a psychology assessment clinic. The subject sample was comprised of individuals referred contractually from school districts and privately by parents. Reasons for subjects' referral to the clinic included poor academic achievement, classroom behavior problems, and suspected intellectual impairments. Fifty-nine percent of the sample were age 5–10 or below (utilizing, by state mandate, a 1 standard deviation IQ/achievement discrepancy for LD a determination) and 41% were age 11–18 (utilizing, by state mandate, a 1 1/2 standard deviation IQ/achievement discrepancy for a LD determination).

Examiners. Twenty-nine examiners participated in this research. They were divided into three categorical groups based on level of experience, as follows: (1) Novice (n=6 testers; 12 evaluations). Students in the second year of an EdS program in School Psychology completing a 1-year practicum in assessment. These students had completed a 4-credit hour course in IQ assessment and a 4-credit hour course in psycho-educational assessment. In addition, they had taken 1/2 semester of prepracticum training at the clinic, during which time they were observed administering both the WISC-III and the W/J-R by licensed School Psychologists and evaluated to insure competency. (2) Intermediate (n=16 testers; 77 evaluations). Students engaged in an internship in School Psychology. These students had completed the same coursework as the novice testers and a full year of school psychology practicum. (3) Advanced (n=7 testers; 20 evaluations). Paid employees of the testing center who were licensed as School Psychologists or Psychologists.

Scorers. One of three individuals rescored each WISC-III and W/J-R protocol. These three individuals were employed by the clinic to ensure scoring consistency and accuracy. As a component of the job requirement, these three individuals completed a supervised training program involving instruction and practice in test scoring. At the time of this study, these individuals had each scored approximately 50 test protocols.

Instruments

The WISC-III (Wechsler, 1991). The WISC-III is comprised of 13 subtests, grouped into three categories: Verbal intelligence (VIQ), Nonverbal intelligence (PIQ), and Full Scale intelligence (FSIQ). The following subtests were evaluated: Information, Similarities, Arithmetic, Vocabulary, Comprehension, Digit Span, Picture Completion, Coding, Picture Arrangement, Block Design, and Object Assembly.

The W/J-R. The W/J-R battery (Woodcock & Johnson, 1989) is comprised of 18 subtests. The new W/J-III (Woodcock, McGrew, & Mather, 2001) has recently been released. Although the entire W/J-R was administered, the present study evaluated only the W/J-R tests that are unchanged in the new version: Letter–Word Identification, Passage Comprehension, Word Attack, Calculation, Applied Problems, and Writing Samples.

Procedures

WISC-III and W/J-R protocols evaluated in this study were initially administered to each subject and scored by one of the 29 examiners. The scorers were given test protocols completed and scored by the initial examiners, which contained information transcribed verbatim during the evaluation. Protocols were randomly assigned to one of the three secondary scorers. When discrepancies were found, these were scrutinized and further evaluated by supervisors, if necessary, until it was determined whether the discrepancy between the original and rescored protocol was an error on the part of the original examiner. Discrepancies that could easily be verified as errors (e.g., an error of addition or an error in transcribing numbers) were not subjected to further verification. Errors of a more subjective nature (e.g., errors in scoring the Writing Samples) were evaluated by a supervisor.

RESULTS

Interrater Reliability

The relationship between the original and rescored tests was evaluated using Pearson product–moment correlations for each subtest on the WISC-III and W/J-R. A Bonferroni adjustment was made due to the use of multiple tests, and alpha levels were set at a value of 0.002 for all pairs of subtest coefficients. Preliminary analysis of the data revealed two data points that substantially affected the results (a scoring error of 41 points on the Letter–Word Identification subtest and a scoring error of 58 on Passage Comprehension); therefore, these outliers were removed for this analysis only. Interrater Pearson reliability coefficients of 0.97–1.00 were obtained for all subtests with the exception of Writing Samples, which received an $r = 0.95$.

Effect of Scoring Errors on Broad Scores

Mean change scores were computed to evaluate the amount of discrepancy (error) from the original scores to the rechecked scores. Mean error scores ranged from a high of 2.8 on Writing Samples to mean error scores of 0 on Digit Span and Word Attack. However, mean scores may be misleading as the majority of change scores were 0, with a few instances of very large error

(e.g., an error of 58 on Passage Comprehension, 41 on Letter–Word Identification, 22 on Writing Samples). Therefore, modal value of scoring errors may be the more appropriate statistic. Table 1 displays the number of protocols containing errors on each subtest and the modal value of the error on each subtest. As can be seen, two of the IQ subtests (Vocabulary and Comprehension) had errors on approximately 1/4 of the protocols; however, these errors were generally quite small (modal value=1). The effect of the scoring errors appears to be minimal, as the mean change in the VIQ, PIQ, and FSIQ was 1.04, 0.8, and 0.76, respectively. On the achievement test, there were large numbers of errors on the Writing Samples (errors on 38 of 91 protocols) subtests. These errors resulted in a mean change of 1.75 points on the Broad Written Language Cluster. The Reading and Math Cluster scores were relatively unchanged by subtest scoring errors.

Table 1
Number of Errors on IQ and Achievement Subtests

Test	Number of protocols with errors (total *n*)	Modal change
IQ test scores		
Information	11 (93)	1
Similarities	12 (93)	1
Arithmetic	3 (93)	1, 3, 6 (one each)
Vocabulary	21 (93)	1
Comprehension	26 (93)	1
Digit Span	0 (81)	0
Picture Completion	3 (92)	1
Coding	1 (91)	1
Picture Arrangement	5 (92)	1
Block Design	3 (93)	3
Object Assembly	1 (87)	1
VIQ	35 (95)	1.04[a]
PIQ	14 (95)	0.28
FSIQ	34 (95)	0.76[a]
Achievement test scores		
Letter Word Identification	7 (84)	2
Passage Comprehension	2 (84)	2, 66 (one each)
Broad Reading Cluster	7 (83)	0.16[a]
Word Attack	0 (61)	0
Basic Reading Cluster	3 (61)	0.008[a]
Calculation	10 (88)	2
Applied Problems	3 (90)	3
Broad Math Cluster	10 (88)	0.25[a]
Dictation	20 (93)	2
Writing Samples	38 (91)	6
Broad Written Language Cluster	38 (91)	1.75[a]

[a]For cluster and broad areas, mean score, rather than mode, is presented.

Expertise of Examiners

Two MANOVAs (using the General Linear Model method for unequal sample sizes) were performed, one for the IQ scores and one for the achievement scores. The independent variable, expertise, had three levels: novice, intermediate, and advanced. The dependent variable was the change score, calculated as the absolute value of the subtest score obtained by the original examiner compared to the score obtained by the second scorer. For the IQ test scores, the multivariate test was significant using Wilks' Lambda criterion [$F(11,63)=1.76$, $P<0.03$; observed power = 0.95], indicating that there was a significant difference in rescored tests, based on level of expertise of the tester. Univariate post hoc tests showed differences by level of expertise on the following subtests: Similarities [$F(2,90)=3.12$, $P<0.04$, observed power = 0.62]; Arithmetic [$F(2,90)=4.12$, $P<0.02$, observed power = 0.57]; and Comprehension [$F(2,90)=7.15$, $P<0.001$, observed power = 0.86]. In all instances, more experienced testers had lower error scores. For the achievement test scores, there were no differences between original and second test scores, by level of expertise of tester. This indicates that although Writing Samples had a high number of scoring errors, these were made across all levels of testers.

Effect of Scoring Errors on Learning Disability Determination

The third statistical procedure was conducted to evaluate the effect of incorrectly scored protocols on subject LD qualification. The McNemar test was performed to evaluate the significance of change in LD qualification based on rescored protocols. Based on the initial test scores, 40 students qualified for a LD diagnosis and 64 students did not qualify for a LD diagnosis. A nonsignificant degree of change was associated with scoring errors. Rescored protocols accounted for only two cases of change in status of student LD qualification among the sample of 110 evaluations. These changes consisted of one failure of a previously LD-qualified individual to obtain LD status, and one qualification of a previously disqualified individual to obtain LD services. Both of these changes in LD classification were on protocols completed by an "intermediate" tester.

DISCUSSION

IQ/achievement discrepancy is not the sole determination of a learning disability. However, it continues to be used in many states as part of the decision making process. Therefore, issues of scoring accuracy are important. The magnitude of score change and interrater disagreement was most prominent for the WISC-III Comprehension and Vocabulary subtests. These results suggest pervasively greater difficulty in reaching interrater

agreement in scoring verbal items. Although there were errors on almost 25% of Vocabulary and Comprehension protocols, the magnitude of errors tended to be small, with an average change in the Verbal IQ score of one point. Vocabulary was the only subtest score with overall interrater reliability falling below 0.97. On the W/J-R, higher interrater disagreement was observed in scoring patterns involving the Writing Samples subtest. Correspondingly, this subtest appears to be the primary contributors towards errors on the Broad Written Language Cluster, which contained an average change score of 1.75 points after rescoring.

Student trainees in school psychology, including practicum students and interns, spend a great deal of time conducting educational evaluations that affect student placement decisions. It is important to verify whether lack of experience can result in scoring errors that might affect placement decisions. Some supervisors/agencies handle this possibility by rechecking all psycho-educational tests for scoring accuracy. Is this precaution necessary? Are trainees susceptible to significant scoring errors on common IQ and achievement tests? What impact do typical scoring errors have on overall evaluation results? The present study suggests that, as expected, level of scoring errors decreases as experience increases. Most notably, trainees were significantly more likely than experienced school psychologists to make scoring errors on the Similarities, Arithmetic, and Comprehension subtests of the WISC-III. Interestingly, on the achievement subtest most susceptible to scoring errors, Writing Samples, there were no differences found across level of expertise. Rather, this subtest tended to present scoring difficulties for all testers.

For the most part, although there were numerous scoring errors found, they tended to be of small magnitude and to have a negligible impact on overall evaluation results. As noted earlier, the Verbal IQ score, on average, changed by one point after rechecking, while the Broad Written Language Cluster changed by almost two points. No other broad band or cluster scores were significantly affected by scoring errors. Of interest was whether learning disability determination would be affected by changes in IQ or achievement scoring errors. It was found that, of 104 cases, two learning disability determinations would have changed as the result of scoring errors. Although this result was not statistically significant, a decision that affects whether or not a child receives exceptional student services is clinically significant. Conventions about the permissible size of the probability of making an error might be modified given the consequences of decision making, e.g., in cases where an error might result in rejecting a qualified child for much needed special services.

Overall, for both the WISC-III and the W/J-R, results of this study corroborate previous research findings of strong interrater reliability. The fact that this study utilized a population of student participants to administer and score these two instruments did not appear to adversely affect

their stability, contrary to prior hypotheses. Additionally, ample interrater reliability coefficients were observed pertaining to individual subtests, as well as broad IQ and area scores.

A new version of the Woodcock/Johnson Achievement test (W/J-III, 2001) has recently been published; therefore, this study only evaluated subtests unchanged in the new version. An evaluation of the new W/J-III subscales would be advisable: These include (on the standard battery) Reading Fluency, Story Recall, Understanding Directions, Math Fluency, Spelling, and Writing Fluency. The 2001 manual now includes a Writing Evaluation Scale to assist examiners in evaluating longer written passages (Mather & Woodcock, 2001). The effect of this additional scoring help should be evaluated. It is hoped that this might bring the interrater reliability of this subscale in line with the high reliabilities found among the other subtests.

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.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders (4th ed., Text Revision)*. Washington, DC: Author.

Cuenot, R. G., & Darbes, A. (1982). A comparison of interscorer agreement for the comprehension, similarities, and vocabulary subtests of the WISC and WISC-R. *Educational and Psychological Measurement, 42*, 417–421.

Dumont, R., Willis, J., & McBride, G. (2001). Yes, Virginia, there is a severe discrepancy clause, but is it too much ado about something? *The School Psychologist, 55*, 1–15.

Education of All Handicapped Children Act of 1975, 20 U.S.C. Sec. 401 (1975).

Individuals with Disabilities Act, 20 U.S.C. Sec. 1400 (1990).

Individuals with Disability Education Act Amendments of 1997, Public Law 105-17, 34 CFR Part 300 (1999).

Mather, N., & Woodcock, R. W. (2001). *Examiner's manual: Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock–Johnson III*. Itasca, IL: Riverside Publishing.

Park, G. E. (1999). Competency-based WISC-III administration based on error frequencies. Doctoral dissertation, Fuller Theological Seminary School of Psychology, 1999. Dissertation Abstracts International, 60 (6-A), A1906. Rehabilitation Act of 1973, 20 U.S.C. Sec. 794.

Shinn, M. R., Good, R. L., & Parker, C. (1999). Noncategorical special education services with students with severe academic deficits. In D. J. Reschly, W. D. Tilly, & J. P. Grimes (Eds.), *Special education in transition: functional assessment and noncategorical programming* (pp. 81–106). Longmont, CO: Sopris West.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. In D. Wechsler (Ed.), *Wechsler Intelligence Scale for Children-Third Edition* (pp. 420–428). San Antonio, TX: Psychological Corporation.

Simpson, R. G., & Halpin, G. (1995). Psychometric effects of altering the ceiling criterion on the Passage Comprehension test of the Woodcock–Johnson Psycho-Educational Battery-Revised. *Educational and Psychological Measurement, 55*, 630–636.

Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: a national survey. *Journal of Psychoeducational Assessment, 12*, 331–350.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition*. San Antonio, TX: Psychological Corporation.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.