Scoring Reliability on the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III)

Joseph J. Ryan
Summer D. Schnakenberg-Ott
Central Missouri State University

Nineteen psychologists and 19 graduate students scored two Wechsler Adult Intelligence Scale–Third Edition patient protocols. Mean IQs and indexes were similar across groups, but the ranges for Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) on one protocol were 25, 22, and 11 points, respectively. For both protocols taken together, percentages of agreement with the "actual" IQs for psychologists were only 26.3 for VIQ, 36.8 for PIQ, and 42.1 for FSIQ. For students, percentages were 15.8 for VIQ, 23.7 for PIQ, and 31.6 for FSIQ. The percentages of FSIQs that fell within ± 1 standard error of measurement of the actual IQs were 89.5 for psychologists and 76.3 for students. Scoring error also had a negative impact on index scores. Both groups were confident about their scoring accuracy.

Keywords: WAIS-III; scoring reliability; indexes; scoring error; confidence rating

It is well known that psychological examiners, regardless of experience level, make numerous errors when scoring the Wechsler scales of adult intelligence. Franklin, Stillman, Burpeau, and Sabers (1982) had certified school psychologists and school psychology students administer the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1955) to one of four specially prepared clients. Each client memorized a script of responses as well as a group of specific behavioral characteristics that he or she would display during the administration (e.g., quiet and withdrawn or hostile and highly verbal). Examination of the completed protocols revealed numerous administration and scoring problems. Typical errors involved improper discontinuance of subtests and failure to properly credit individual responses on the Information, Comprehension, and Vocabulary subtests. Scoring and administration errors were even noted on the Digit Span and Digit Symbol subtests. Some examiners discontinued Digit Span prior to failure on both trials of an item, whereas others continued testing even though the termination criteria had been met. On Digit Symbol, some examiners were careless in their scoring and assigned credit for one or more incorrect number-symbol pairings. Franklin et al. (1982) noted that the magnitude of error introduced by poor administration and scoring procedures could easily result in misplacement or exclusion of individuals from special programs and/or produce invalid test results.

Ryan, Prifitera, and Powers (1983) used the Wechsler Adult Intelligence Scale–Revised (WAIS-R) (Wechsler, 1981) and compared the scoring accuracy of 19 Ph.D. psychologists with an average of 7.3 years of testing experience with that of twenty 2nd-year psychology graduate students. Each participant scored WAIS-R protocols from one male and one female vocational rehabilitation client. Actual protocols were used, as opposed to fictitious ones containing ambiguous responses, because test records containing many ambiguous responses may not accurately reflect the manner in which real clients perform on the WAIS-R. Results indicated that both seasoned practitioners and inexperienced graduate students made numerous scoring errors that produced marked variability in obtained IQs. Examination of the protocols revealed errors such as incorrectly converting sums of scaled scores to IQs, giving too much or too little credit to individual items,

Correspondence concerning this article should be addressed to Joseph J. Ryan, Department of Psychology, 1111 Lovinger Building, Central Missouri State University, Warrensburg, MO 64093; e-mail: ryan@cmsu1.cmsu.edu.

Assessment, Volume 10, No. 2, June 2003 151-159
DOI: 10.1177/1073191103252348

and calculation mistakes when adding raw scores of subtests. Moreover, psychologists had significantly greater variability than did the students on the Performance IQs (PIQs) of both protocols and were more likely to make errors when determining the IQ than were the students. Psychologists produced PIQs that ranged from 119 to 129 on one protocol (the actual IQ was 122) and 88 to 105 on the other (the actual IQ was 99). Students generated PIQs that ranged from 122 to 126 on Protocol 1 and 98 to 102 on Protocol 2. For both protocols taken together, the proportions of participants who calculated IQs within ± 1 standard error of measurement (*SEM*) of the actual IQs were 88.5% on the Verbal Scale, 94.5% on the Performance Scale, and 82.5% on the Full Scale. The *SEM* is used to estimate the amount of variability in an individual's score. However, this statistic does not include the impact of inadequate scoring on test accuracy. Thus, clerical and mechanical problems in scoring constitute error over and above the known chance variability associated with a test score (Kaufman, 1990).

Three investigations focused exclusively on the performance of master's level graduate students enrolled in psychological assessment courses. Slate and Jones (1990a) inspected 149 student-generated protocols and found that novice examiners experienced difficulty assigning correct point values to verbal responses. This problem contributed to overestimation or underestimation of Full Scale IQs in 56% (range = 1 to 10 IQ points) and 16% (range = 1 to 2 IQ points) of the protocols, respectively. For 12% (n = 18) of the protocols, the overestimates were 4 or more IQ points. In a second study, Slate and Jones (1990b) analyzed 180 WAIS-R protocols from 26 student examiners and found approximately nine errors (SD = 5.6) per protocol, with 98% of the protocols having at least one error. Typical scoring problems included incorrect point assignments to individual items, failure to credit items below the basal level, and assigning credit to items above the ceiling. In both investigations, the highest frequency of scoring mistakes occurred on the Vocabulary, Comprehension, and Similarities subtests.

Slate, Jones, and Murray (1991) also evaluated the effect of testing practice on the administration and scoring proficiencies of 20 graduate students. They found that students who administered the WAIS-R on five occasions made more errors on the fifth examination than they did on the first. Moreover, administrative proficiencies did not improve even after 10 examinations. With respect to scoring errors, the authors noted that students scored protocols incorrectly because (a) they did not understand the scoring criteria provided in the test manual, and (b) they tended to make mistakes due to carelessness. Lack of understanding of scoring criteria was evidenced by failure to properly credit responses to the Vocabulary, Comprehension, and Similarities subtests, whereas carelessness was indicated by inadequate recording of responses, incorrectly converting raw scores to scaled scores or IQs, and frequent computational errors.

Now that the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III) (Wechsler, 1997) has been published, clinicians will incorporate this instrument into their assessment batteries. Until proven otherwise, they will likely assume that the impact of scoring errors on the WAIS-III has been reduced, or is at least unchanged, from that reported for previous editions of the scale. This assumption may not be accurate since the WAIS-III is a more complex instrument than is either the WAIS or WAIS-R. The new scale usually requires the examiner to score 13 subtests and to calculate four index scores, three IQs, and two supplementary measures of incidental memory. Conversely, the WAIS and WAIS-R require only the scoring of 11 subtests and three IQs.

The present study was designed to examine scoring reliability on the WAIS-III using two separate protocols. The first (Protocol 1) was obtained from a 62-year-old man with a high school education and a clinical diagnosis of organic brain syndrome. The second (Protocol 2) was from a 36-year-old woman with 12 years of education who sustained a mild head injury approximately 20 months prior to testing. Each protocol contained the exact responses of the examinees as well as response times and other data necessary for scoring 13 subtests and obtaining the three IQs, four indexes, and two incidental memory scores. A second purpose of the study was to investigate whether persons with differing levels of training and experience differ with respect to scoring variability on the WAIS-III. To accomplish this goal, we used a group of doctoral-level psychologists who regularly conducted intellectual evaluations and a group of graduate students who had recently completed formal training in the administration and scoring of the WAIS-III.

METHOD

Participants

Twenty-five doctoral-level psychologists with extensive testing experience were mailed two WAIS-III protocols and asked to score them and then indicate their degree of confidence in the accuracy of the results. Each psychologist was contacted to ensure her or his interest and co operation prior to mailing the materials. A package containing the two protocols, a cover letter, a questionnaire (requesting information on years of testing experience subsequent to the terminal degree, number of WAIS-III administrations, and overall confidence in the accuracy of

their scoring), a research consent form, and a return envelope was mailed to each psychologist. The first author conveyed to the participants that they were to score each protocol entirely by hand.

Twenty-five graduate students who had recently completed a course in individual intelligence testing were contacted and asked to take part in the study. Each was provided with a package containing the two protocols, a cover letter, a questionnaire (requesting information concerning the number of WAIS-III administrations completed and a rating of overall confidence in the accuracy of their scoring), and a research consent form. The student participants were asked to return the scored protocols and completed questionnaire within 2 weeks to the senior author. Participants were directed to score each protocol entirely by hand.

Procedure

From the files of the first author, protocols of one male and one female patient were randomly selected. Both patients had been referred for a comprehensive neuropsychological evaluation. The man, who was an inpatient at a midwestern Veterans Affairs medical center, was referred by a staff neurologist and carried a working diagnosis of organic brain syndrome with possible seizure disorder and possible dementia. His electroencephalograph was consistent with bitemporal dysfunction and a magnetic resonance imaging study indicated cortical atrophy. The female was a private-practice referral to the first author from a legal nurse consultant for a major law firm. She sustained a mild head injury without loss of consciousness during a motor vehicle accident. She was transported to a community hospital for emergency treatment but discharged approximately 2 hours later with a diagnosis of generalized muscle pain/strain and instructions for coping with head injury, back pain, and strain. She had a 3 cm raised bump on the left forehead that was treated by application of an ice pack. On a follow-up visit to a private physician, she complained of depression, confusion, absentmindedness, and irritability and was given a diagnosis of mild traumatic brain injury with postconcussion syndrome.

The first author and a consultant independently scored each protocol. The consultant has extensive experience with the Wechsler scales of intelligence and regularly teaches a graduate course that covers administration and scoring of the WAIS-III. Next, a meeting of the first author and the consultant was held in order to achieve 100% agreement on the scoring of each protocol via item-by-item reviews. When there were disagreements on specific items, the appropriate sections in the *WAIS-III Administration and Scoring Manual* (Wechsler, 1997) were consulted (i.e., scoring responses on pp. 45-50 and the sample items for the individual subtest), and the raters discussed their rationale for the assigned scores. Discussion of individual items continued until consensus was reached for each disputed response. Disagreements primarily involved point assignments (i.e., 0, 1, or 2 points) for responses to the Vocabulary, Similarities, and Comprehension subtests.

The resulting indexes, IQs, and incidental memory scores were designated as "actual" scores. Each returned protocol was checked for accuracy by the authors, and all errors (e.g., computational and clerical mistakes) were recorded. Means, standard deviations, and ranges for the indexes, IQs, and incidental memory scores were obtained separately for the psychologists and graduate students and then compared across groups. Participants were also asked to indicate how confident they were in their scoring of each protocol using separate 7-point Likert-type scales (*not confident* = 1, *confident* = 4, and *extremely confident* = 7). Means, standard deviations, and ranges for the confidence ratings were calculated separately for the psychologists and graduate students and then assessed for possible group differences.

RESULTS

Of the 25 psychologists who agreed to participate in the study, 20 returned the scored protocols and completed the confidence ratings. However, only 19 were usable. The 20th protocol had been scored using *SAWS: A Scoring Assistant for the Wechsler Scales for Adults* (The Psychological Corporation, 1997). Because the purpose of the study was to evaluate all aspects of WAIS-III scoring accuracy, this protocol was eliminated from the data set. Fifteen of the individuals who provided usable protocols held a Ph.D., 3 were Psy.D.s, and 1 was an Ed.D. Testing experience of the psychologists averaged 11.92 years (*SD* = 8.19). All 19 participants were actively engaged in assessment practice within institutions and/or private practice settings. The median and mean numbers of total WAIS-III administrations were 22 and 54.89 (*SD* = 102.08), respectively. Five individuals had not administered the WAIS-III because they used technicians and/or advanced graduate students for this purpose. Nevertheless, these participants regularly evaluated the scoring accuracy of their subordinates and felt qualified to participate in the present study. Fifteen of the participants had experience within Department of Veterans Affairs medical centers as psychologists, predoctoral interns, and/or postdoctoral fellows, and 3 regularly taught graduate courses in individual intelligence testing. Thus, it was reasonable to assume that all partici-

TABLE 1
Means, Standard Deviations, Ranges, *SEM*s, and Confidence Limits for Actual IQs

	Protocol 1						*Protocol 2*					
	VIQ		*PIQ*		*FSIQ*		*VIQ*		*PIQ*		*FSIQ*	
Scoring Group	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ph.D.s (n = 19)	90.42 (87-92)	1.57	92.42 (90-100)	2.85	91.10 (87-97)	2.38	98.42 (94-106)	2.38	91.78 (91-97)	1.36	96.16 (94-100)	1.42
Students (n = 19)	91.57 (88-99)	2.17	91.53 (79-94)	3.51	91.05 (86-97)	2.12	99.00 (88-113)	5.07	93.05 (87-109)	5.17	96.79 (89-110)	4.30
Actual IQs	92		91		91		99		91		96	
*SEM*s	2.35		3.27		2.07		2.47		3.54		2.23	
Confidence limits for actual IQs	90-94		88-94		89-93		97-101		87-95		94-98	

NOTE: Ranges of actual IQs are in parentheses. *SEM* = standard error of measurement; VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full Scale IQ.

pants had previous experience scoring WAIS-III protocols and also had experience evaluating patients similar to those in the present study.

Eighteen of the original 25 graduate students returned scored protocols and completed confidence-rating forms. To have an equal number of participants in both groups, a 19th graduate student volunteer was subsequently located and recruited for the study. This individual had recently completed a graduate course in individual intelligence testing. The median and mean numbers of total WAIS-III administrations were 5 and 5.42 (SD = 1.02), respectively.

Table 1 reports means, standard deviations, and ranges for the IQs obtained by the psychologists and graduate students on both protocols. Also reported are the "actual" IQs earned by the two examinees along with the associated *SEM*s and confidence limits. Although the IQ means are similar across scoring groups for each protocol, the standard deviations and ranges provide clear evidence that scoring error had a meaningful impact on the accuracy of the WAIS-III IQs. For instance, psychologists had a range of 5 points on the VIQ and 10 points on both the PIQ and FSIQ of Protocol 1. Students produced ranges of 11 points on the VIQ, 15 points on the PIQ, and 11 points on the FSIQ. For Protocol 2, the ranges for psychologists were 12 points on the VIQ and 6 points on the PIQ and FSIQ. Students produced ranges of 25 points on the VIQ, 22 points on the PIQ, and 21 points on the FSIQ. In every instance, the student-generated ranges were larger than those calculated for the psychologists. For both protocols taken together, the percentages of perfect agreement with the "actual" IQs for psychologists were 26.32 for VIQ, 36.84 for PIQ, and 42.11 for FSIQ. For students, the corresponding percentages were 15.89 for VIQ, 23.68 for PIQ, and 31.58 for FSIQ. The percentages of psychologists' scores that fell within ± 1 *SEM* of the actual IQs for both protocols taken together were 76.32 for VIQ, 92.11 for PIQ, and 84.21 for FSIQ. Corresponding percentages for students were 65.78 for VIQ, 89.47 for PIQ, and 73.67 for FSIQ.

Table 2 reports means, standard deviations, and ranges for the indexes obtained by the psychologists and graduate students on both protocols. Also reported are the "actual" indexes achieved by the two examinees along with the associated *SEM*s and confidence limits. Inspection of the ranges indicates that on Protocol 1 the Perceptual Organization Index (POI) varied from 4 points, when scored by psychologists, to 19 points, when scored by students. On Protocol 2, the Verbal Comprehension Index (VCI) range for psychologists was 4 points, whereas the VCI range for students was 15 points. For both protocols taken together, the percentages of perfect agreement with the "actual" indexes for the psychologists were 36.84 for VCI, 60.53 for POI, 73.68 for Working Memory Index (WMI), and 65.78 for Processing Speed Index (PSI). For students, the corresponding percentages were 21.1 for VCI, 31.6 for POI, 81.6 for WMI, and 78.9 for PSI. The percentages of psychologists' scores that fell within ± 1 *SEM* of the actual indexes for both protocols taken together were 73.68 for VCI, 92.11 for POI, 89.47 for WMI, and 94.74 for PSI. The corresponding percentages for students were 81.57 for VCI, 92.11 for POI, 92.11 for WMI, and 97.37 for PSI.

The scores generated by the students and psychologists did not differ for either protocol on any of the IQ and index means. However, when the variances were compared for Protocol 1, the students demonstrated significantly greater scoring variability than the psychologists on the POI, $F(2, 18) = 9.91$, $p < .01$. On Protocol 2, students demonstrated significantly greater scoring variability than psychologists on the VIQ, $F(2, 18) = 4.54$, $p < .01$; PIQ, $F(2, 18) = 14.53$, $p < .002$; FSIQ, $F(18, 18) = 9.15$, $p < .01$; and VCI, $F(2, 18) = 8.99$, $p < .01$. Students demonstrated significantly less variability than psychologists on the PSI, $F(2, 18) = 5.03$, $p < .01$.

TABLE 2
Means, Standard Deviations, Ranges, *SEM*s, and Confidence Limits for Actual Indexes

	Protocol 1								*Protocol 2*							
	VCI		*POI*		*WMI*		*PSI*		*VCI*		*POI*		*WMI*		*PSI*	
Scoring Group	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ph.D.s (n = 19)	86.21	2.90	101.74	1.66	93.32	1.95	81.42	1.80	102.57	1.26	92.89	1.73	79.26	3.12	92.57	2.71
	(80-88)		(99-103)		(88-95)		(79-88)		(101-105)		(88-93)		(69-82)		(91-103)	
Students	87.74	2.16	101.05	5.23	92.84	2.69	80.52	2.87	104.16	3.78	93.37	1.74	79.84	2.36	91.37	1.21
(n = 19)	(86-94)		(84-103)		(84-94)		(69-84)		(101-116)		(88-95)		(71-84)		(88-93)	
Actual indexes	88		101		94		81		105		93		80		91	
*SEM*s	2.77		3.36		3.71		4.83		3.02		3.75		3.87		4.91	
Confidence limits for actual indexes	85-91		98-104		90-98		76-86		102-108		89-97		76-84		86-96	

NOTE: Ranges of actual indexes are in parentheses. *SEM* = standard error of measurement; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; WMI = Working Memory Index; PSI = Processing Speed Index.

A number of subtests were particularly difficult to score. For Protocol 1, psychologists experienced the most difficulty with Vocabulary followed by Picture Completion, Comprehension, and Similarities. These subtests were also the most difficult to score for graduate students. Considering Protocol 2, psychologists encountered difficulty arriving at correct scores, in descending order, for Comprehension, Digit Symbol-Coding, Vocabulary, and Similarities. The Comprehension subtest was also the most difficult to score for the graduate students, followed by Digit Symbol-Coding, Vocabulary, and Similarities. Inspection of the individual protocols indicated that scoring variability resulted from an assortment of mistakes and that none of the protocols were error-free. The mean numbers of errors for psychologists and students on Protocol 1 were 9.79 (SD = 3.60) and 10.21 (SD = 4.20), respectively. For Protocol 2, the means were 7.21 (SD = 2.76) for psychologists and 8.53 (SD = 2.74) for the students. The two scoring groups did not differ significantly in the number of errors they produced on Protocol 1, $t(17) < 1$, or Protocol 2, $t(17) = 1.42$, $p > .15$.

The impact of scoring error on WAIS-III interpretation was evaluated by determining the number of times the FSIQs of Protocols 1 and 2 (a) yielded ability ranges that differed from those of the actual IQs and (b) fell outside two *SEM*s (i.e., ± 4 points) of the actual IQs. Ability ranges were based on Wechsler's (1997) seven qualitative descriptors for WAIS-III FSIQ scores. For Protocol 1, the actual FSIQ was 91, a value falling within the average range and at the 27th percentile rank. On two occasions, the IQ values generated by psychologists dropped one classification, from average to the low average. One participant calculated a FSIQ of 87 (19th percentile rank), whereas the second produced an IQ of 88 (21st percentile). In both cases, the IQ values fell within ± 4 points (i.e., 87 to 95) of the actual FSIQ. There was one instance in which a psychologist obtained an IQ of 97. This summary value fell within the average ability classification but was beyond the upper limit of the selected confidence range. On Protocol 2, which had an actual FSIQ of 96 (average range and 39th percentile rank), there were no differences between the ability classifications of the psychologists and that of the actual FSIQ. All FSIQs fell within the specified confidence limits of 92 to 100.

For Protocol 1, two of the students generated IQs that fell within the low average range, one classification below that of the actual FSIQ. One student calculated an IQ of 89 (23rd percentile rank), and the other obtained an IQ of 86 (18th percentile). The former score fell within the specified confidence limits (i.e., 87 to 95) of the actual FSIQ of 91, whereas the latter score fell outside the confidence limits. A third student obtained an IQ of 97, a value falling outside the designated confidence limits but within the average ability classification. On Protocol 2, which had an actual FSIQ of 96, one student obtained a high average IQ of 110 (75th percentile rank), another obtained an average IQ of 102 (55th percentile rank), and a third calculated a low average IQ of 89 (23rd percentile rank). Two of the IQs represented ability classifications different from that of the actual FSIQ, and all three summary scores fell beyond the specified confidence limits (i.e., 92 to 100).

The scoring confidence rating means of the psychologists were 5.18 (SD = 1.10) and 5.10 (SD = 1.30), respectively, for Protocols 1 and 2. Students' average scoring confidence rating for Protocol 1 was 4.68 (SD = 1.39), and for Protocol 2, the mean confidence rating was 4.84 (SD = 1.63). The confidence ratings did not differ reliably between students and psychologists for Protocol 1, $t(36)$ =

1.23, $p < .15$, or Protocol 2, $t(36) < 1$. To investigate intrascorer reliability and potential relationships between scoring accuracy and scoring confidence, a series of Pearson product-moment correlation coefficients were calculated. To control the experiment-wise error rate, the Bonferroni correction was applied (.05/10 = .005). The correlation between numbers of errors in Protocol 1 and the numbers of errors in Protocol 2 for psychologists was nonsignificant, $r(17) = .077$, as were correlations between the numbers of errors and the confidence ratings for Protocol 1, $r(17) = –.267$, and Protocol 2, $r(17) = –.044$. For students, the correlation between errors in Protocol 1 and Protocol 2 was nonsignificant, $r(17) = .266$, as were the correlations between the numbers of errors and degree of scoring confidence for Protocol 1, $r(17) = .244$, and Protocol 2, $r(17) = .456$.

Nonsignificant correlations were found between the number of scoring errors in the protocols and the amount of WAIS-III administrative experience of the participants. The correlations for psychologists were $r(17) = –.319$ on Protocol 1 and $r(17) = .007$ on Protocol 2. For students, the correlations were $r(17) = .030$ for Protocol 1 and $r(17) = –.064$ for Protocol 2.

DISCUSSION

The findings of this investigation are consistent with the literature (e.g., Franklin et al., 1982; Ryan et al., 1983; Slate & Jones, 1990b) because they indicate that regardless of one's experience level with the Wechsler scales, scoring errors occur frequently and detract from the accuracy of WAIS-III IQs and indexes. However, in the study by Ryan et al., psychologists and students achieved similar levels of scoring accuracy on the Verbal Scale but differed significantly in terms of scoring variability on the Performance Scale, with psychologists demonstrating significantly more variability than students. In the present study, the findings were reversed because the students' scoring variability was significantly greater than that of the psychologists on one or more summary components of both protocols. Perhaps sampling differences are partially responsible for the difference. In the Ryan et al. study, each student possessed a master's degree and was enrolled in a clinical psychology doctoral program at an urban university. In the present investigation, the graduate students were each working toward a terminal master's degree at a university in the rural Midwest and were required only to have completed a course in individual intelligence testing. This is, of course, a tentative explanation of scoring differences between psychologists and students because, as mentioned above, previous research and the present findings (e.g., nonsignificant correlations between number of scoring errors and amount of WAIS-III administrative experience) suggest that an examiner's experience level is not of critical importance when it comes to scoring accuracy on the Wechsler scales (Kaufman, 1990). However, it may be that there is an experience threshold (e.g., a course in individual intelligence and administration and scoring of a large number of supervised examinations) that must be reached before the impact of this variable washes out.

The present findings indicate that both psychologists and students demonstrated considerable variability in scoring the WAIS-III. This is particularly troublesome because both groups were confident that they had scored the protocols accurately. Students and psychologists had average confidence ratings of 4.68 (Protocol 1) and 4.84 (Protocol 2) and 5.18 (Protocol 1) and 5.10 (Protocol 2), respectively. A rating of 4.0 on the Likert-type scale indicated a confident examiner. Moreover, the participants were volunteers who had completed formal training in individual intelligence testing. Therefore, it was assumed that they (a) were motivated to do their best on the scoring task and (b) knew exactly what was expected of them. If these assumptions are accurate, the present results are consistent with Kaufman's (1990) assertion that scoring errors are an unfortunate built-in aspect of individual assessment.

For both psychologist and student participants, there were no meaningful associations between scoring accuracy (i.e., numbers of errors in a protocol) on Protocol 1 and Protocol 2. Thus, intrascorer reliability was lacking, and it was not possible to identify participants who were either consistently good or consistently bad at scoring the protocols. Perhaps the results would have been more encouraging had we employed a greater number of participants and required them to score a larger sample of protocols. Another noteworthy finding was the lack of association between scoring confidence and scoring accuracy for psychologists and students on both protocols. The fact that participants were confident about their scoring and simultaneously error prone suggests a number of possibilities. Perhaps they dealt with answers that were not clearly scorable using the test manual by "reading into" the responses information that was not present. This might cause an examiner to assign too much credit to one or more responses from the Comprehension or Vocabulary subtests. Another possibility is that they were unaware of their errors because of a failure to double-check each protocol for correct scoring (Slate &Hunnicutt, 1988).

The *WAIS-III Administration and Scoring Manual* (Wechsler, 1997) represents a significant improvement over its predecessors, the WAIS and WAIS-R, because it presents expanded instructions for examiners along with increased numbers of examples on how to score individual items. Nevertheless, it appears that these improvements

had little impact on the accuracy of scores produced by the present samples of students and practitioners. Perhaps scoring problems on the WAIS-III reflect the same difficulties that were identified by Slate et al. (1991) when they evaluated WAIS-R scoring accuracy. That is, in addition to simple carelessness, many examiners my not have fully understood the scoring criteria provided in the manual. To compensate for the latter possibility, specialized teaching techniques could be developed to improve the scoring reliability of student examiners. This might involve the use of special classroom scoring exercises, programmed workbooks, and/or videotaped WAIS-III administrations that require students to record and score an examinee's responses. For both students and psychologists, the development of a detailed WAIS-III scoring supplement might also prove valuable. Massey, Sattler, and Andres (1978) published a scoring supplement for the Wechsler Intelligence Scale for Children–Revised (WISC-R) (Wechsler, 1974) that was widely used by students and practitioners. No empirical investigations of the effectiveness of the supplement have been published. However, anecdotal reports from students and practitioners who used this tool in conjunction with the WISC-R manual indicated that it helped them achieve greater scoring precision.

Consistent with previous research, the Vocabulary, Comprehension, Similarities, and Digit Symbol-Coding subtests were among the most difficult to score for both students and practitioners (e.g., Franklin et al., 1982; Slate et al., 1991; Slate & Jones, 1990a, 1990b). Conversely, the present study yielded some unexpected results as well because the Picture Completion subtest posed scoring problems for the present participants. Thus, 89.5% (17/19) and 68.4% of students and practitioners, respectively, incorrectly scored Item 10 (i.e., leaf) on Protocol 1. In the vast majority of cases, credit was given to a spoiled response that involved pointing correctly while simultaneously giving a verbal response that was clearly incorrect. This problem appears to reflect, at least in some cases, either a disregard for scoring instructions presented on page 67 of the administration manual or a lack of understanding concerning a spoiled response. If and when a supplemental scoring guide for the WAIS-III is developed, the future authors should seriously consider including a detailed explanation for, and numerous examples of, spoiled responses for individual subtests.

In the present study, every protocol contained one or more errors. For instance, psychologists and students failed to give credit for unadministered items above the basal and/or assigned too much or too little credit to individual items that were passed. Likewise, credit was sometimes given to failed or spoiled items but withheld on correctly answered items. In one instance, a 2-point credit was given to each item above the basal when the correct value was actually 1 point each. Additional problems included adding points incorrectly, using a supplementary subtest when determining the IQ, and not subtracting for obvious errors on Digit Symbol-Coding. Overall, errors that reflected conceptual difficulties (i.e., whether to assign a response a 0-, 1-, or 2-point credit) applying the standard scoring criteria to individual items were far more prevalent than those due to simple carelessness in scoring (e.g., adding points incorrectly or using the wrong table to convert raw scores to scaled scores). This observation suggests that the use of a computerized scoring system (e.g., SAWS) will not eliminate the majority of errors that detract from the reliability of subtest, index, and IQ scores.

The Digit Symbol-Incidental Learning procedure was scored correctly by 18 of the 19 psychologists. However, one participant assigned credit to an incorrect recollection on the pairing component of Protocol 1. Eighteen students scored Incidental Learning correctly, but 1 participant added the pairing items incorrectly and failed to record any of the freely recalled symbols. Examination of both protocols indicated that this student counted the correct pairs in row 1 of the Response Booklet and then entered this number as the total pairing score on page 4 of the protocol. The participant then scored the second row of the pairing component and entered this number on page 4 of the protocol as the free recall score. With respect to the Digit Symbol-Copy procedure, one psychologist calculated a score of 53, but the correct number of symbols copied was actually 43. One student made a similar error by recording a score of 33 when 43 symbols had actually been copied. Minor scoring errors occurred for two psychologists and three students because of a problem with the Response Booklet for Digit Symbol-Copy. In the second row (Item 14) of the Response Booklet, the symbol to be copied is a three-sided U-shaped figure open to the left. However, the scoring template for this item presents a figure that is open at the top. If examiners use the template to score the item, a correctly copied symbol will be scored as an error. Sattler and Ryan (2001) noted this problem and recommended that when scoring Item 14, examiners disregard what is shown on the scoring template and give credit for a drawing that matches the model.

Finally, three practical implications of the present study need to be considered. First, the degree of unreliability reported above addresses only the issue of scoring precision. Problems with poor test administration (Moon, Blakey, Gorsuch, & Fantuzzo, 1991), the examinee's physical and emotional state during testing (Hanna, Bradley, & Holen, 1981), and examiner-examinee characteristics (Slate & Hunnicutt, 1988) were not studied. Because so many uncontrolled variables contribute to reduced-test reliability, it is essential to report individual WAIS-III IQs and indexes in conjunction with a confidence interval based on

either the *SEM* or standard error of estimate (*SEE*), whichever is appropriate to the testing situation (Sattler, 2001). Of course, confidence intervals based on these statistics provide conservative estimates of the unreliability associated with a given test score because they are based on internal consistency coefficients and account only for content sampling error. Perhaps the test-retest stability coefficient or some other measure of reliability should be used to calculate the *SEM* and *SEE* and determine confidence limits for the Wechsler IQs. This would yield larger confidence limits than have been reported previously for the WAIS-III and underscore the fact that considerable uncertainty/error is associated with an examinee's obtained IQ.

A second practical implication has to do with the negative consequences of imprecise scoring on everyday practice. When frequent errors occur in a test protocol, this may alter the summary scores to a point where the intelligence classification is incorrect and the client is either denied needed services or placed in an inappropriate work, school, or treatment situation. The fact that the students calculated FSIQs for Protocol 2 that varied by as much as 21 points suggests that erroneous placement based on imprecise scores is a distinct possibility. When FSIQs in the present study were interpreted, the ability classifications changed from average (i.e., 90-109) to low average (i.e., 80-89) in five instances and from average to high average (i.e., 110-119) in one instance. The protocols used in this investigation were obtained from persons with average intelligence, and it might be argued that a change from the average to low-average ability classification would have little or no impact on the examinee. This may be the case for persons with average intellectual ability, but a change of one classification could be potentially problematic for individuals with greater or lesser intellectual ability. Academic decisions and training opportunities might change for someone with superior (i.e., 120-129) intelligence if he or she received an erroneous FSIQ in the high-average range. Likewise, an examinee with mild mental retardation (i.e., IQ = 50-55 to approximately 70; American Psychiatric Association, 2000) might receive inappropriate placement if he or she was incorrectly classified as having borderline intelligence (i.e., 70-79). It would be informative if future research on WAIS-III scoring reliability used protocols from individuals at both extremes of the IQ distribution. Perhaps it is easier to score responses from persons with mental retardation than it is to score answers provided by individuals with average to superior intelligence.

Finally, in cases of traumatic brain injury, the WAIS-III is often administered along with specialized neuropsychological measures to estimate the extent of the patient's behavioral and cognitive impairments. When personal injury litigation is initiated, the WAIS-III becomes part of the legal record. Under these circumstances, a protocol that contains numerous errors may damage the credibility of the examiner, hurt the reputation of his or her profession, and have a real impact on the outcome of the proceedings.

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

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

Hanna, G., Bradley, F., & Holen, M. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology*, *19*, 370-376.

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

Massey, J. O., Sattler, J. M., & Andres, J. (1978). *WISC-R: Scoring supplement for the Wechsler Intelligence Scale for Children, Revised Edition*. Palo Alto, CA: Consulting Psychologists Press.

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

The Psychological Corporation. (1997). *SAWS-A: Scoring assistant for the Wechsler scales for adults*. San Antonio, TX: Author.

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

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.

Sattler, J. M., & Ryan, J. J. (2001). WAIS-III subtests and interpreting the WAIS-III. In J. M. Sattler, *Assessment of children: Cognitive applications* (4th ed., pp. 415-454). San Diego, CA: Author.

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

Slate, J. R., & Jones, C. H. (1990a). Examiner errors on the WAIS-R: A source of concern. *Journal of Psychology*, *124*, 343-345.

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

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*, 375-379.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children–Revised.* New York: The Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Joseph J. Ryan is a professor of psychology and chair of the Department of Psychology at Central Missouri State University (CMSU). Prior to joining the faculty at CMSU, he was chief, Psychology Service at the Dwight D. Eisenhower Veterans Affairs Medical Center in Leavenworth, Kansas. He received his Ph.D.

from the University of Missouri–Columbia and is a diplomate in clinical neuropsychology of the American Board of Professional Psychology (ABPP). His general research interest are in neuropsychological assessment.

Summer D. Schnakenberg-Ott is a 1st-year doctoral student in clinical psychology at the Forest Institute of Professional Psychology in Springfield, Missouri. She received her master's degree in psychology from Central Missouri State University in December 2001. Her current research interests are neuropsychological assessment and the detection of malingered performance on the Wechsler Adult Intelligence Scale–Third Edition.