Graduate Student WAIS-III Scoring Accuracy Is a Function of Full Scale IQ and Complexity of Examiner Tasks

Christopher J. Hopwood
Texas A&M University

David C. S. Richard
Rollins College and Eastern Michigan University

Research on the Wechsler Adult Intelligence Scale–Revised and Wechsler Adult Intelligence Scale–Third Edition (WAIS-III) suggests that practicing clinical psychologists and graduate students make item-level scoring errors that affect IQ, index, and subtest scores. Studies have been limited in that Full-Scale IQ (FSIQ) and examiner administration, recording, and scoring tasks have not been systematically varied. In this study, graduate student participants score a high (FSIQ = 112) and low (FSIQ = 85) IQ record form in one of two stimulus conditions: digitized film clips (N = 13) or partially completed record forms (N = 11). Results demonstrate that examiners are less accurate in the high IQ condition, and that recording examinee responses from scoring video clips results in more scoring errors. Obtained FSIQs are significantly higher than criterion IQ scores in the high IQ condition (8.46 for video condition, 2.55 for record form condition). Self-reported proficiency in WAIS-III administration and scoring is positively related to number of scoring errors.

Keywords: Wechsler scales; scoring errors; clinical decision making; scoring confidence

An intelligence testing course is offered by virtually every psychology department that provides clinical training (Oakland & Zimmerman, 1986), and the Wechsler scales are among the most widely used intellectual assessment tests. Prior surveys of clinical training programs found that the Wechsler Adult Intelligence Scale–Revised was the most frequently used measure of adult intelligence and that training in its administration occurred in 88% of intelligence testing courses (Slate, Jones, & Murray, 1991). We surveyed 27 American Psychological Association–accredited graduate programs via phone interview. All 27 reported providing training in the Wechsler scales and the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III). Eighteen programs (67%) reported that Wechsler training was emphasized as much as or more than any other intelligence test.

The rich history of the Wechsler scales has been marked by concerns regarding the interrater reliability of some subtests (Ryan, Prifitera, & Powers, 1983). Interrater reliability is the degree to which raters or independent observers agree on the dimensions (e.g., occurrence, magnitude) of an event or person being measured (Haynes & O'Brien, 2000). With regard to the Wechsler scales, concerns about interrater reliability have focused primarily on verbal subtests that require significant examiner judgment in scoring (e.g., vocabulary, similarities, and comprehension) and the digit symbol coding subtest. Scoring errors that either overestimate or underestimate a person's true IQ impair the validity of clinical inferences made from test results.

The issue of interrater reliability has been studied in two ways. One method has been to report measures of agreement (e.g., Pearson Product-Moment correlations, intraclass correlations) between two or more raters across subtests and scales. For example, the *WAIS-III Technical Manual* (The Psychological Corporation, 1997) reports a

Assessment, Volume 12, No. 4, December 2005 445-454
DOI: 10.1177/1073191105281072

study in which 60 record forms (RFs) from the original standardization sample were randomly selected and independently scored by three raters. Raters were required to score transcribed items, calculate raw scores, and transform raw scores into scaled scores. They did not record item responses or administer the test. Intraclass correlation coefficients for vocabulary, similarities, and comprehension were .95, .93, and .91, respectively. The range of obtained subtest raw or scaled scores was not reported. A similar study was recently reported in the Wechsler Intelligence Scale for Children, fourth edition (WISC-IV; Wechsler, 2003). Sixty cases were sampled from the WISC-IV standardization sample and scored by four doctoral-level raters. Intraclass correlations were .98 for similarities, .98 for vocabulary, .95 for comprehension, .96 for information, and .97 for word reasoning.

An alternative method is to assess rater accuracy by comparing item scores provided by raters to an a priori criterion score determined by investigators. Thus, disagreement (or low interrater reliability) represents the degree to which participant scoring decisions deviate from the a priori criterion. Because accuracy studies rely less on correlational data, include a greater number of raters, and report results in more descriptive terms, they provide a picture that is more complex than interrater reliability studies reporting intraclass correlations. For example, Ryan and Schnakenberg-Ott (2003) recently had 19 psychologists and 19 graduate students score two WAIS-III protocols that were partially completed (i.e., item responses were included but scores were not). Practicing clinicians obtained perfect agreement with criterion IQ scores on 26.3% of verbal IQ (VIQ) scores, 36.8% of performance IQ (PIQ) scores, and 42.1% of Full Scale IQ (FSIQ) scores. Students obtained perfect agreement with the criterion on 15.8% of VIQ scores, 23.7% of PIQ scores, and 31.6% of FSIQ scores. Although the mean obtained FSIQs were within 1 or 2 points for both groups on the two protocols, the range of obtained FSIQs was large across protocols and groups. FSIQ ranges in the Ph.D. sample were 10 points for Protocol 1 and 6 points for Protocol 2; student sample FSIQ ranges were 11 points for Protocol 1 and 21 points for Protocol 2. These results were generally consistent with previous findings regarding scoring accuracy on the Wechsler scales (e.g., Franklin, Stillman, Burpeau, & Sabers, 1982; Ryan et al., 1983; Slate & Jones, 1990a; Slate, Jones, Murray, & Coulter, 1992). However, the finding that practicing clinicians tended to outperform students in obtaining perfect agreement with criteria represents a departure from previous reports that have found that students tend to commit fewer errors than professionals (e.g., Ryan et al., 1983; Slate et al., 1992) and that practice is not associated with a decrease in scoring and administration errors (Slate et al., 1991; Patterson, Slate, Jones, & Steger, 1995).

Across studies, some errors have been found to be more common than others. The three most common errors have been failure to record examinee responses, assignment of incorrect points, and failure to query appropriately. When failure to record responses is excluded from the errors, vocabulary, comprehension, and similarities evidence the most examiner errors largely because of the complexity of the scoring rules and reliance on examiner judgment for these subtests (Patterson et al., 1995; Slate & Jones, 1990b; Slate et al., 1991; Slate et al., 1992).

Limitations of Previous Studies

The methodology of previous research has been limited to studying the extent to which scoring errors influence accuracy of obtained FSIQs. As Ryan and Schnakenberg-Ott (2003) note, error can also be introduced in a variety of other ways (e.g., by virtue of poor test administration, the examinee's physical or mental condition, examiner-examinee characteristics). Because professionals and students were only presented with partially completed protocols, any error associated with administration practices was removed from experimental consideration. Therefore, results from previous studies may underestimate examiner errors given the exclusion of sources of error variance associated with administration practices. A goal of our study was to manipulate examiner task complexity by requiring examiners to either score partially completed protocols or record responses from digitized video clips (VID) of a WAIS-III administration. Our first hypothesis was that individuals required to both record participant responses and score the protocol would make significantly more errors than individuals scoring partially completed protocols. Rejection of the null hypothesis would suggest that prior studies estimating error rates by using partially completed protocols represent an underestimate of true error.

We also hypothesized that the number of errors on high IQ protocols would be significantly greater than the number of errors on low IQ protocols and that both the range and variances of obtained FSIQs would be greatest in the high IQ condition. We reasoned that protocols with high IQ scores might be associated with more errors because they are associated with more non-0-point responses and thus more opportunities to commit scoring errors. We chose a design that provides additional information beyond descriptive data characteristic of previous studies to examine the effect criterion IQ and stimulus format have on rates of scoring errors.

METHOD

Participants

This study was approved by the Human Subjects Review Board at Eastern Michigan University; all participants provided informed consent prior to participation. Participants were master's and doctoral students in clinical psychology who successfully completed a semester-long course in Wechsler intelligence testing within the past year. Coursework involved a minimum of five WAIS-III and WISC-III administrations with case reports. In addition, students were either videotaped or directly observed on two occasions and provided feedback regarding their proficiency at administering and scoring the test.

Participants were matched on a composite variable comprising self-reported undergraduate grade point average (GPA),[1] intelligence testing course instructor, and orientation of graduate program track in the Eastern Michigan Graduate Program (traditional clinical vs. behavioral) and assigned to either the VID or RF condition, as described below. One case from the RF group was removed from the analysis because of error scores greater than 3 standard deviations above the mean on both RFs. Thus, 13 participants were retained for the VID condition and 11 for the RF condition. Twenty-one of the 24 participants were female; data regarding age was not collected.

In the VID condition, participants reported an average undergraduate GPA of 3.54 ($SD = 0.39$); the mean GPA for RF participants was 3.66 ($SD = 0.34$). Participants also reported the number of Wechsler administrations they had conducted in the past year and completed four Likert-type self-efficacy items that assessed the examiner's perception of their own WAIS-III proficiency, perceived usefulness of the test, interest in intellectual assessment, and general personal competence. These data were used for several analyses described below.

Stimulus Materials

Two criterion protocols were designed to obtain FSIQ scores approximately 1 *SD* below and above the mean. Items from the protocols of several individuals in the second author's files were selected for the criterion protocols. Thus, each protocol was a composite of several actual cases. Item responses were selected based on their score; secondary consideration was given to the degree of ambiguity of the item with respect to manual exemplars (less ambiguous item responses were favored over more ambiguous item responses). Item responses were extracted from the protocols to construct criterion RFs that were then independently scored by the authors. Disagreements about item scoring were resolved by discussion. The items were then reevaluated by the first author and a rater blind to experimental hypotheses to confirm correct item scores. The final FSIQ scores for the high and low protocols were 85 and 112, respectively.

Design and Procedure

The study employed a 2×2 mixed ANOVA design. The between-subjects factor was stimulus format and the within-subjects factor was IQ of a simulated case. With regard to stimulus format, participants in the RF condition received a partially completed protocol with transcribed, typewritten item responses and times to completion for Performance scale items. Participants were required to complete item scores, sum subtest raw scores, and calculate all relevant standard scores (i.e., IQs and index scores). Participants in the VID condition received a blank protocol and viewed digitized VID of an actress scripted to provide the same responses as those provided in the protocol for the RF condition. The actress responded to off-camera cue cards with the scripted response for each item (including cues regarding speed to item completion on performance tasks). The VIDs were digitized and presented to participants in Microsoft PowerPoint. Each new slide represented an item and contained the appropriate VID, an item identifier (e.g., "Picture Completion, Item 1"), an option to see the clip again, and an option to move to the next clip. The option to see the clip again was included to correct for any clips that may have been difficult to hear or for which the participant was not ready; no data were recorded regarding its use. Actresses in both the high IQ and low IQ conditions were female research assistants presented to participants as being 22 years old.

The within-subjects factor was the level of scripted FSIQ of the simulated case (FSIQ = 85 or 112). Thus, each participant scored two protocols with true FSIQs of 85 and 112 in either the RF or VID condition. Participants completed the high and low IQ conditions in counterbalanced order and were instructed to score every item on both RFs and ignore reversal and discontinue rules because they implied scoring decisions.

IQs, indexes, and subtest-scaled scores were analyzed in terms of deviation from the criteria, deviation between groups, and variability. The primary dependent variable for inferential statistics was scoring errors, defined as deviations from the criterion RF. Errors were categorized according to four classifications: item-level scoring errors, clerical errors, mathematical calculation errors, and item-level timing errors. Item-level scoring errors were deviations from item scores on the criterion RF. An item-level timing error was one in which an incorrect score was recorded for a timed item (these errors were not counted as item-level scoring errors). A clerical error involved any in-

accuracy that occurred in transcribing or scaling subtest raw scores. A mathematical error was an error in summing item raw scores or subtest-scaled scores. A research assistant blind to the design and hypotheses was trained to identify and calculate errors for each completed RF. Errors were then double checked by the first author for accuracy.

RESULTS

The term *criterion* is used to refer to the relevant consensually determined IQ, index, subtest-scaled score, or item raw score for each protocol. The term *stimulus* refers to the way the information was presented to the participant: either in VID format or as a partially completed RF. The term *obtained score* refers to the entry recorded by the participant.

Error Analysis

A total of 418 ($M = 8.71$, $SD = 3.54$) errors were identified across the 48 protocols, of which the 381 item-level scoring errors ($M = 7.94$, $SD = 3.23$) accounted for 91% of all errors made. Participants also made 25 mathematical calculation errors across RFs ($M = 0.52$, $SD = 0.85$), 7 timing errors ($M = 0.15$, $SD = 0.55$), and 5 clerical errors ($M = 0.10$, $SD = 0.37$). The sum of all errors was the dependent variable for statistical analyses.

Equality of variance for the high and low IQ conditions was assessed using the Pitman test (Pitman, 1939). Error scores in the high IQ condition were significantly more variable than those in the low IQ condition, $r(22) = .864$, $p < .001$. A test for homogeneity of variance of error scores in the between-group factor found no statistically significant difference between the VID and RF conditions. According to Stevens (2002), parametric tests such as ANOVA are appropriate for within-group comparisons of means despite unequal variances when the variance ratio is less than 4:1 and sample sizes are not disproportionate, so a 2×2 mixed ANOVA was selected to test mean group differences. Participants made more errors on the high IQ ($M = 9.67$, $SD = 3.64$) than low IQ protocol ($M = 7.75$, $SD = 3.22$). This difference was statistically significant, $F(1, 22) = 6.37$, $p < .05$, $d = .56$, partial $\eta^2 = .22$, and represented a moderate effect size (Cohen, 1969). There was also a large between-subjects main effect (Cohen, 1969) in which participants in the VID condition made significantly more errors ($M = 9.88$, $SD = 3.67$) than participants in the RF condition ($M = 7.32$, $SD = 2.87$), $F(1, 22) = 5.44$, $p < .05$, $d = .78$, partial $\eta^2 = .20$. Although the mean number of errors was greatest in the high IQ/VID cell ($M = 11.15$, $SD = 3.53$), the IQ $\times$ Stimulus condition interaction was not significant. Together, the two main effects suggest that errors increased both as a function of criterion IQ and the stimulus condition. The high IQ profile resulted in significantly more scoring errors independent of the number of items scored, because participants scored every item on every subtest (i.e., the discontinuation rules were not employed). The VID condition, and presumably the increased tasks associated with it, also resulted in significantly more errors than the RF condition.

Although most errors related to scoring mistakes, mathematical and clerical errors may have a greater potential to dramatically alter FSIQ scores. For example, an item-level scoring error on the vocabulary subtest may change the raw score by up to 2 points. This may in turn change the scaled score by a maximum of 1 point, which could change the FSIQ by a maximum of 1 point. However, mathematical (e.g., miscalculating the sum item scores to derive a scale raw score) and clerical (e.g., mistranslating a subtest raw score to its correct scaled score) errors could lead to a drastic change in the FSIQ. In our data, 25 mathematical errors were committed. Nine of these had no potential effect on the FSIQ (e.g., mathematical errors on symbol search or letter-number sequencing subtests affect index scores but not IQ scores). However, the other 16 mathematical errors affected FSIQ, on average, by 5.5 points. Three of the five observed clerical errors had potential to affect the FSIQ and did so by 1.33 points, on average. Nineteen of the mathematical errors and all five of the clerical errors occurred in the VID condition, raising some concerns about results with respect to primary hypotheses. Thus, we ran the same 2×2 ANOVA described above with item-level scoring errors. Both between-subject, $F(1, 22) = 4.58$, $p < .05$, and within-subject, $F(1, 22) = 4.70$, $p < .05$, effects remained statistically significant, suggesting that mathematical and clerical mistakes alone do not explain findings related to primary hypotheses.

Obtained IQ, Index, and Subtest Scores

Table 1 reports means, standard deviations, and ranges for all obtained IQs and index scores. The table also reports the criterion IQs and index scores from both the low and high IQ protocols (the value to the left is the score for the low IQ protocol). Means for both the high and low IQ conditions were higher than their respective criterion scores (high IQ mean = 117.75, $SD = 8.24$; low IQ mean = 85.71, $SD = 3.43$; see Table 1 for criterion scores). Results for the high IQ condition were statistically significant, $t(23) = 3.418$, $p < .01$, whereas results for the low IQ condition were not, $t(23) = 1.011$, $p = .323$. All but one of the obtained mean FSIQs, VIQs, and PIQs were higher than their respective criterion IQs. The greatest deviation

TABLE 1
Mean Obtained IQs and Index Scores

		VID Condition (n = *13*)						*RF Condition* (n = *11*)					
		Low IQ			*High IQ*			*Low IQ*			*High IQ*		
	Criterion Score Low/High	M	SD	*Range*	M	SD	*Range*	M	SD	*Range*	M	SD	*Range*
IQs													
Full Scale	85/112	86.00	3.83	82 to 97	120.46	9.04	113 to 141	85.36	3.04	82 to 94	114.55	6.11	109 to 132
Verbal	85/115	86.08	2.60	82 to 93	116.62	6.05	111 to 134	86.45	2.34	84 to 93	115.36	6.10	111 to 133
Performance	86/107	86.15	2.15	84 to 91	118.62	8.24	109 to 134	85.82	3.18	83 to 95	111.91	5.07	105 to 124
Index scores													
VCI	86/114	85.92	2.18	80 to 89	115.23	3.32	112 to 124	86.64	1.12	86 to 89	113.00	2.86	109 to 118
POI	88/111	86.69	1.80	84 to 89	110.15	2.82	105 to 116	86.00	1.79	82 to 88	106.45	4.57	99 to 111
WMI	88/113	86.23	5.29	69 to 90	112.54	0.88	111 to 113	88.55	1.57	86 to 90	115.02	0.45	109 to 115
PSI	88/99	90.15	4.56	86 to 99	101.62	3.60	99 to 106	88.0	0.0	0	100.09	1.87	99 to 103

NOTE: VID = video clips; RF = record form; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; WMI = Working Memory Index; PSI = Processing Speed Index.

TABLE 2
Obtained Subtest-Scaled Scores

		VID Condition (n = *13*)				*RF Condition* (n = *11*)			
		Low IQ		*High IQ*		*Low IQ*		*High IQ*	
	Criterion Score Low/High	M	SD	M	SD	M	SD	M	SD
Verbal scale									
Vocabulary	9/13	9.2	.56	13.4	0.77	9.4	.67	12.9	0.83
Similarities	7/12	7.2	.56	12.2	1.2	7.0	.0	11.6	0.93
Information	6/13	5.7	.86	13.0	0.0	6.0	.0	13.0	0.0
Comprehension	8/13	8.38	.65	13.4	1.20	8.0	.63	12.6	0.52
Arithmetic	8/12	8.1	.28	11.8	0.38	8.1	.30	12.0	0.45
Digit span	7/12	7.0	.0	11.9	0.28	7.1	.30	11.8	0.60
Letter-number	9/13	8.54	.88	13.0	0.0	9.18	.87	13.0	0.0
Performance scale									
Block design	8/12	8.0	.71	11.62	0.87	8.0	.0	12.0	0.0
Matrix reason	8/12	8.0	.0	12.1	0.28	8.0	.0	12.0	0.0
Picture composition	8/12	7.38	.77	12.0	0.41	7.0	.89	10.2	1.9
Picture arrangement	8/12	8.0	.0	13.7	2.0	8.0	.0	12.3	0.91
Symbol search	8/8	8.15	.90	8.46	0.97	8.0	.0	8.27	0.47
Coding	8/12	8.6	1.5	12.0	0.0	8.0	.0	12.0	0.0

NOTE: VID = video clips; RF = record form.

involved the mean FSIQ in the VID/high IQ condition (M = 120.46, SD = 9.04), which was 8.46 points (more than half a standard deviation) higher than the actual FSIQ of 112.

Assuming raters represent a sample drawn from a population of infinite raters, the standard deviation of obtained standard scores (e.g., IQ, index, or subtest) within each group represents an empirical estimate of the standard error of measurement (SEM) interpretable in IQ units. Table 1 demonstrates that higher IQ scores and increased tasks both generally resulted in higher SEM estimates. For example, with respect to FSIQ, the largest SEM occurred in the VID/high IQ condition, followed by the RF/high IQ, VID/low IQ, and RF/low IQ conditions. All four were larger than the FSIQ SEM provided in the *WAIS-III Technical Manual* (The Psychological Corporation, 1997) for individuals ages 20 to 24 (SEM = 2.37) and for the SEM averaged across all age groups (SEM = 2.30). Results were somewhat mixed for VIQ and PIQ as well as index scores. As can be seen in Table 1, the VIQ and PIQ SEMs are roughly the same across VID and RF groups but are greater for high as opposed to low IQ scores.

The largest range in obtained FSIQ scores occurred in the high IQ/VID condition (i.e., 28 points). In fact, Table 1 shows that the ranges for the high IQ/VID condition were consistently greater than the other conditions. The only exception to this trend was for the Working Memory Index and Perceptual Organization Index, in which the high IQ\RF conditions evidenced slightly greater ranges. Six of the possible eight index score comparisons also showed greater range in the high IQ than low IQ condition.

We ran t tests to examine between-group differences in obtained IQ scores. For the low IQ protocol, none of the obtained IQs or index scores were significantly different as a function of stimulus condition. For the high IQ protocol, the obtained mean PIQ in the VID condition was significantly higher than that in the RF condition, $t(22) = 2.35$, $p < .05$. However, this was likely due to several participants having committed mathematical errors on the scaled score table that artificially inflated PIQ scores rather than item-level scoring mistakes inflating the PIQ. This finding highlights the dramatic effect of this relatively infrequent type of error.

Table 2 shows obtained subtest-scaled scores. Highest rates of agreement with the criterion protocol for subtest-scaled scores were found for the matrix reasoning (97.9%), digit symbol-coding (95.8%), and information (95.8%) subtests. Consistent with previous research, the lowest rates of agreement and highest rates of item-level scoring errors were found for the vocabulary (25.0%, 10.35%), comprehension (54.2%, 11.07%), and similarities (64.6%, 6.77%) subtests. Vocabulary, comprehension, and similarities also accounted for 79.01% of item-level scoring errors made across all subtests (39.86%, 23.82%, and 12.2%, respectively).

Agreement With Criteria

The percentage of perfect agreement between obtained FSIQs and the criterion FSIQs was greater in the low (41.7%) than in the high (4.2%) IQ condition. Perfect agreement on the VIQ score was 16.7% for low IQ proto-

cols and 8.3% for high IQ protocols; on the PIQ, it was 29.2% for low IQ protocols and 4.2% for high IQ protocols. Perfect agreement with index criteria varied in both high (Verbal Comprehension Index [VCI] = 16.7%, Perceptual Organization Index [POI] = 45.8%, Working Memory Index [WMI] = 75.0%, Processing Speed Index [PSI] = 66.7%) and low (VCI = 66.7%, POI = 37.5%, WMI = 50.0%, PSI = 83.3%) conditions.

Classification agreement, or the percentage of participants who correctly classified the low IQ individual as low average across all IQ and index scores and the high IQ individual as high average across FSIQ, VIQ, VCI, POI, and WMI and average with respect to PIQ and PSI was also assessed. Classification agreement was lower for the high IQ condition than the low IQ condition and was particularly low for the PIQ (20.83%) and POI (50.00%) on high IQ RFs. Excluding these percentages, the IQ and index scores were correctly classified at rates ranging from 70.83% to 100.00%, with a mean of 90.63%. Although lower agreement for the high IQ condition is a trend consistent with analyses regarding perfect agreement, the findings for the PIQ and POI appear to be the result of criterion scores that were close to the cutoff of classification categories. For example, although 50% of participants misclassified high IQ protocols in terms of the POI, the criterion score of 111 was 1 point from the classification cutoff.

Experience, Self-Rated WAIS-III Proficiency, and Error Scores

Prior to having participants score the RFs, we asked them to report the number of Wechsler protocols they had administered in the previous year. Participants reported an average of 7.63 Wechsler administrations ($SD = 4.1$). Errors were negatively but not significantly correlated to the number of administrations ($r = -.13$, $p = .38$) and to undergraduate GPA ($r = -.15$, $p = .30$), and no statistically significant differences were observed between conditions in terms of experience. We also asked participants to rate their own proficiency on the WAIS-III relative to their peers, how useful they perceived the test to be, their interest in intellectual assessment, and how much they may worry about not functioning well as a psychologist. Ratings were made on a 7-point Likert-type scale, with 1 = *not at all*, 4 = *moderate*, and 7 = *very much so*. No statistically significant mean differences were observed between conditions on any of these items. Overall, participants largely perceived themselves as being proficient at administering and scoring a WAIS-III relative to their peers ($M = 5.17$, $SD = 0.92$). We examined whether self-reported proficiency in WAIS-III scoring and administration was related to number of errors made. Within-group error scores for the high and low IQ conditions were significantly correlated, $r(22) = .46$, $p < .01$. Thus, we summed each participant's total error score across IQ conditions and correlated it with self-reported scoring proficiency. Somewhat unexpectedly, self-reported scoring proficiency correlated positively with the number of errors committed, $r(22) = .300$, $p < .05$. Thus, the greater a participant's self-rated proficiency in WAIS-III administration and scoring, the more errors they made.

Ambiguity of Item Responses

To assess the level of item response ambiguity in criterion RFs, we classified responses as either ambiguous or not ambiguous. An ambiguous response is not directly referenced in the scoring exemplars provided in the manual. An unambiguous response is directly referenced as an exemplar in the WAIS-III manual. Item responses on relevant subtests were classified as ambiguous or not ambiguous by a psychologist blind to study hypotheses, and classifications were confirmed by the first author. In all, 92.9% of items that required some subjective judgment in scoring (i.e., picture completion, vocabulary, similarities, comprehension, and information) had direct scoring referents in the WAIS-III manual.

To test whether response ambiguity had an effect on scoring accuracy, ambiguous and unambiguous 0-point responses were selected from the vocabulary subtest of the low IQ protocol. The item response sample was selected because it included the highest frequency of ambiguous and unambiguous responses (seven of each) of any subtest with a three-tiered item scoring format. We constrained our ambiguity analysis to 0-point responses on the low IQ RF to control for response complexity when making our comparison and because, again, it provided the most data. We then compared the accuracy of obtained item scores for the seven ambiguous and unambiguous vocabulary responses across participants. Participants across both VID and RF conditions made a total of 10 scoring errors on unambiguous items and 8 on ambiguous items. These results do not support the hypothesis that the ambiguity of item responses mediated number of scoring errors made. However, this conclusion should be viewed tentatively because of the post hoc nature of the analysis and the limited data available to address the question.

DISCUSSION

To investigate the effects of the complexity of examiner tasks and FSIQ on WAIS-III scoring accuracy, we asked graduate students to score protocols with predetermined IQ scores of 85 and 112 in one of two conditions: partially completed RFs or digitized film clips of examinee re-

sponses. Participants committed an average of 8.71 scoring errors per protocol. The range of obtained FSIQs was 32 points in the high IQ condition, consistent with prior research demonstrating wide ranges in obtained FSIQs when compared to a criterion RF. Scoring errors compromised the ability of participants to achieve satisfactory rates of classification and score agreement with criterion RFs across all IQ and index scores, particularly in the high IQ condition.

Our first hypothesis—that errors would increase as a function of stimulus condition—was supported. Participants in the VID group made significantly more errors than participants in the RF group. Errors appear to be a function of the number and type of administration and scoring tasks required of the examiner. IQs were calculated incorrectly more often, and the SEM increased in the VID group relative to the RF group, indicating both that transcribing examinee responses onto RFs appears to increase the total number of errors made and that errors tended to be nonrandom (i.e., tended to increase obtained scores). It could be that the increasing administration demands made participants less vigilant regarding correct IQ calculation practices. Results suggest that future work should explore the impact of all levels of administration and scoring on the reliability of scores.

The hypothesis that participants would make more errors in the high IQ than low IQ conditions was also supported. At face value, this hypothesis is intuitive: More opportunities to err result in more errors. However, this result is important in particular because the SEM increased in the high IQ condition. As efforts were made to generate item responses of similar scoring difficulty across IQ score conditions and discontinuation rules were not employed, neither item difficulty nor fatigue appears to explain this result. One possible explanation is that there were more opportunities to err on non-0 responses, as is the case with progressively higher IQ scores. Because errors were not random but instead increased with IQ scores, the SEM for the higher FSIQ was greater. In contrast to our findings, confidence intervals (CIs) assume uniform error across the distribution of IQ scores. For example, the CI widths for FSIQ scores of 85 (95% CI = 82 to 89) and 112 (95% CI = 108 to 115) are both 7 points. Our results suggest a reconsideration of the assumption that CIs should be invariant across the continuum of IQ scores.

Our data indicate that individual clerical and mathematical errors had a greater impact on IQ and index scores than did the more common item-level scoring errors. This finding indicates the importance of double checking calculations and of using computer software in scoring the WAIS-III. We also found that mathematical and clerical errors were more common in the VID condition. As the VID condition took considerably longer and was more consistent with an actual administration in terms of time, this finding suggests that examiner fatigue may play a role in WAIS-III scoring errors. However, hypothesized effects remained significant when these types of errors were not included in the analyses, suggesting that the possible relation between fatigue and clerical and mathematical mistakes do not explain our results.

This study replicated the consistent finding that most scoring variability and errors are associated with the vocabulary, similarities, and comprehension subtests. The fact that close to 80% of the item-level scoring errors made by participants were on these three verbal subtests emphasizes the difficulty examiners continue to have with them. This may be due to unclear scoring criteria, a tendency for students to match an examinee response with the first acceptable exemplar in the scoring criteria and overlook better matches worth fewer points, or other cognitive processing strategies worthy of future study. In any case, there was a consistent bias for students to provide too much credit for an examinee response.

Ryan and Schnakenberg-Ott (2003) noted that participants in their sample were generally confident in their ability to score a WAIS-III RF. In their study, correlations between self-reported confidence levels and number of errors made by professionals and students were not significant across protocols. In our study, self-rated proficiency was positively associated with scoring errors. Thus, unlike Ryan and Schnakenberg-Ott's results, we found some evidence that the more confident individuals were in their WAIS-III proficiency, the more errors they made. The discrepancy may be due to the fact that we asked participants to provide us proficiency ratings prior to examining the protocols, whereas Ryan and Schnakenberg-Ott asked participants to rate scoring confidence after they completed the protocols. Thus, ratings in their study reflected a judgment of accuracy on a recently completed task, whereas our ratings were a global self-evaluation of WAIS-III administration and scoring proficiency. Although our results await replication and should be interpreted conservatively given the small sample size, it is fair to conclude that self-reported confidence in WAIS-III proficiency does not positively predict scoring accuracy.

The data did not suggest that having a direct exemplar of item responses in the scoring manual improved scoring accuracy on the vocabulary subtest of the low IQ RF. Instead, participants committed a similar number of errors on ambiguous and unambiguous item responses. However, this finding should be viewed cautiously because this was not a primary study hypothesis and unambiguous item responses were intentionally oversampled. Future research on the effect of item response ambiguity on scoring errors across subtests involving subjective scoring decisions would further clarify this issue.

Our study had several limitations. Most obvious was that participants did not administer the WAIS-III. They were asked to score item-completed RFs or to rate responses that were presented in a Microsoft PowerPoint presentation. Although we reviewed every clip in the VID condition for clarity and did not receive any negative feedback regarding the quality of the digital scenes, the task required of participants differed from an actual testing situation. Video digitization, however, allowed us to control precisely the examinee's responses and ensured identical stimulus presentation to VID participants. A full WAIS-III administration would probably result in higher rates of scoring inaccuracy than we reported given the greater demands of an actual testing situation.

Our study differed from previous efforts (e.g., Ryan et al., 1983; Ryan & Schnakenberg-Ott, 2003) in that we did not compare practicing clinical psychologists to graduate students. Although this sample could be generally characterized by limited experience, within the sample, experience was not significantly correlated with errors, consistent with previous studies (e.g., Slate et al., 1991). In any case, data gathered from a nonprofessional sample may not generalize to the behavior of professionals who regularly administer Wechsler instruments, as it could be that the results regarding scoring inaccuracy reflect limited or unique training experiences. This is especially the case because data from the current study are characterized by higher error rates than in previous studies involving students and professionals (e.g., Ryan & Schnakenberg-Ott, 2003). Conversely, scoring errors may be associated with bad testing habits that are acquired when clinicians become overconfident in their knowledge of the test's administration and scoring rules or are less likely to receive supervisory feedback. Nevertheless, until a similar method is employed in the investigation of scoring practices among professionals, results should not be considered generalizable to standard scoring situations with experienced professionals.

The VID condition took considerably longer than the RF condition, as such differences between the groups in terms of errors may be explained by boredom or fatigue. This is particularly important because the VID condition was more similar than the RF condition to a real-world administration. This finding reemphasizes the more general point that previous studies on scoring accuracy are likely to have considerably underestimated real-world errors.

Our results highlight the importance of referring to the manual before scoring item responses, particularly on vocabulary, comprehension, and similarities subtests. The importance of close manual consultation might also involve a reconsideration of the common practice of scoring during the course of an administration. Although real-time scoring is helpful in deciding when to query and thus may reduce administration errors, it may also increase scoring errors by encouraging examiners to make hasty scoring judgments. Unfortunately, research to date has suggested little with regard to reducing administration and scoring errors (Slate & Jones, 1990b; Slate et al., 1991), and this is an important area of future work. In particular, examiner and examinee characteristics and their interaction need to be investigated to assist the development of training methods that reduce errors. Furthermore, the development of testing materials that permit assessment of important cognitive constructs while minimizing administrative and scoring errors would benefit future versions of the Wechsler scales. To this end, simplified item-level scoring criteria and the use of user-friendly computer algorithms that automatically perform complex scaling calculations may have a significant effect on reducing examiner error.

NOTE

1. Self-reported undergraduate GPA was used rather than graduate GPA because many of the students had completed only one to two semesters of graduate work at the time of the experiment.

REFERENCES

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Haynes, S. N., & O'Brien, W. H. (2000). *Principles and practice of behavioral assessment*. Dordrecht, the Netherlands: Kluwer Academic.

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

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

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, *31*, 9-12.

The Psychological Corporation. (1997). *Wechsler Adult Intelligence Scale–Third Edition technical manual*. San Antonio, TX: Harcourt Brace.

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

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler adult intelligence scale-third edition (WAIS-III). *Assessment*, *10*(2), 151-159.

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

Slate, J. R., & Jones, C. H. (1990b). 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 Wechlser Adult Intelligence Scale–Revised: An empirical evaluation of practice administrations. *Professional Psychology: Research and Practice*, *22*(5), 375-379.

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1992). Evidence that practitioners err in administering and scoring the WAIS-R. *Mea-*

surement and Evaluation in Counseling and Development, *25*, 156-161.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). London: Lawrence Erlbaum.

Wechsler, D. (2003). *WISC-IV Technical and Interpretive Manual*. San Antonio, TX: The Psychological Corporation.

Christopher J. Hopwood is a doctoral candidate in Clinical Psychology at Texas A&M University. His interests include clinical assessment, personality disorders, and interpersonal process.

David C. S. Richard is an associate professor and director of Psychology and Organizational Behavior in the Hamilton Holt School of Rollins College. His research interests include Behavioral assessment, computerized and computer-assisted assessment, accuracy of self-reported behavior, ecological momentary assessment, and the effect scoring and administration errors have on clinical judgment and decision-making.