2006

Conditional error variance in the WISC-IV

Laszlo Attila Erdodi

CONDITIONAL ERROR VARIANCE IN THE WISC-IV

by

Laszlo Attila Erdodi

Thesis

Submitted to the Department of Psychology

Eastern Michigan University

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Clinical-Behavioral Psychology

Thesis Committee:

John Knapp, PhD, Chair

Renee Lajiness-O'Neill, PhD

Dean Lauterbach, PhD

October 17, 2006

Ypsilanti, MI

Acknowledgements

I would like to thank Dr. David Richard for introducing me to the intricacies of intelligence testing, for his encouragement to pursue novel ideas, and for his continued assistance with this project despite the distance.

Also, I am grateful to Dr. John Knapp for restructuring the research design at the proposal meeting and for always being available to share his invaluable statistical expertise with me throughout the thesis-development process.

I would like to acknowledge the special contributions of Christopher Hopwood, Dr. Steven Huprich, and Dr. Carol Freedman-Doan in developing the protocols used in this study.

Finally, I would like to thank Stefan Tolna for preparing the data collection protocols.

Abstract

Measurement error at different ability levels in the WISC-IV was studied to empirically test the conditional error variance hypothesis. Graduate students in clinical psychology at a Midwestern university scored fictitious WISC-IV Vocabulary subtests constructed to yield actual scaled scores of 4, 10, and 16. Classical measurement theory assumes error rate will be constant across the three conditions. Modern test theories (Item Response Theory), however, predict that the precision of a measurement instrument will change as a function of the examinee's ability level. Data supported the conditional error variance hypothesis. Scorers made significantly more errors in the low- and high-ability-level conditions than they did in the average ability condition. Implications of these findings for intelligence testing are discussed.

Table of Contents

List of Tables

List of Figures

Introduction and Background

The Wechsler IQ tests are the most widely used intellectual assessment instruments. Psychologists consider the test scores to be valid indicators of intellectual ability and frequently incorporate them into higher level inferences about the cognitive functioning of tested individuals. The accuracy of IQ scores is important given their significant influence on assessment, treatment, and placement recommendations.

Accuracy is degraded by measurement error, which is inherent in the process of quantifying variables. Measurement error is any variance in obtained scores not due to the variable being measured but still affecting the obtained score (Lord & Novick, 1968). Thus, error is a source of inconsistency, a random or systematic variation that can never be completely eliminated from a measurement instrument.

In classical test theory, error represents the discrepancy between the obtained and the true score (Gregory, 1996). Given that the true score cannot be known, the estimate of error variance yields a confidence interval within which a person's true score is likely to reside. In other words, the precision of a test is defined in terms of the average difference between a hypothetical latent trait level and any given test score. This creates a paradoxical situation in latent variable measurement models, in which an unknown parameter (true score) is estimated on the basis of a number (observed score), the validity of which is derived from its proximity to an unknown parameter. Thus, in the absence of an incontrovertible index, it is critical to have an accurate estimate of the error distribution associated with obtained scores so that they can be correctly interpreted at all ability levels.

The standard error of measurement (SEM) is a descriptive measure of the variability in the error observed in a sample of scores and is a function of both the reliability of a test and the variability in the target population of scores. (Atkinson, 1990). SEM units are given in the same metric as is the test itself. Thus, it also may be thought of as an estimated standard deviation of a hypothetical normal distribution of an individual's obtained scores if the person were to take the same test an infinite number of times. The mean of this distribution would be the person's true score (Aiken, 2000). Although each individual will have a unique true score, in classical measurement theory SEM is conceptualized as a stable property of the test independent of a person's ability level. Because confidence intervals are calculated from the SEM, it follows that the width of the intervals are invariant across all levels of ability as well.

Mathematically, the SEM is most often derived from the population standard deviation of the test score distribution and the reliability coefficient associated with the test. There are several ways to compute the SEM. Each method will produce a slightly different value, and each has unique properties that must be considered before it is applied to a specific data set (Lord, 1984). However, all computational procedures are derived from the basic formula

$$\text{SEM} = \sigma\sqrt{1-r_{xx}}$$

where σ is the population standard deviation and r_{xx} is the selected reliability coefficient. It is apparent that the SEM is proportionate to σ and inversely related to r_{xx}. In other words, measurement error depends on both the actual variability in the measured trait and the error variance (i.e., variability coming from sources extraneous to the target variable). Reliability itself is defined as the ratio of true score variance to observed score variance and cannot be

directly determined given that true scores are never known (Dimitrov, 2002). Consequently, reliability is estimated using statistical and empirical methods (e.g., split-half reliability or Cronbach's alpha, test-retest reliability, criterion-referenced reliability, interrater reliability). Based on the estimated population standard deviation and the chosen reliability coefficient of the test, the SEM is invariant within a given measurement model and represents the first step in calculating the confidence interval within which a person's true score may be found. As previously stated, in classical test theory, the SEM does not change as a function of the true score; therefore, it is treated as a constant unaffected by the natural variability observed in test scores (see Figure 1).

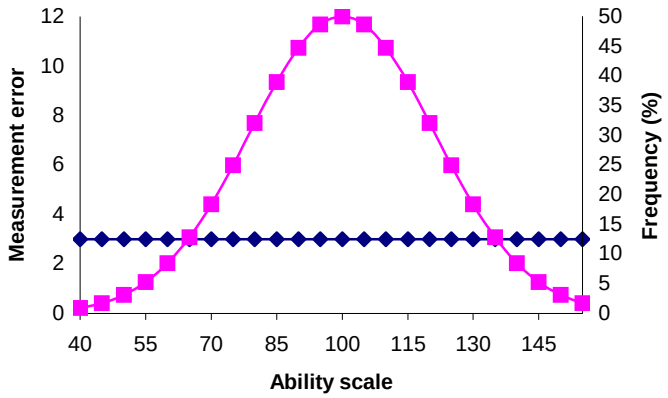

Figure 1. SEM in classical measurement theory applied to the Wechsler intelligence scales (SEM = $\sigma\sqrt{1-r_{xx}}$).

In fact, neither local between-subjects variability nor reliability coefficients at different ability levels (e.g. low, average, high) are constant along the scale of measurement. They also fluctuate as a function of other demographic variables (e.g., type and severity of

coexisting disorders), creating the need for establishing more specific indices of variability and reliability. For example, the WISC-IV manual gives different values for both σ and r_{xx} based on age, subtest, and ability level (Wechsler, 2003), implicitly recognizing that the construct of intelligence may not be equally stable in different segments of the population. Thus, the major disadvantage of using one generic SEM to describe the accuracy of an obtained score is that its actual value may vary along the measurement scale, and no universally accepted methods exist for calculating SEMs specific to each level of ability (Murphy & Davidshofer, 1998).

Another fact that compromises the validity of the SEM in quantifying measurement error is the obvious, yet often ignored, discrepancy between internal consistency and empirical measures of reliability in applied settings. Friedman (1970) studied the long-term (17-month) stability of IQ scores as measured by the WISC in a sample of 44 children in the borderline-low average ability range. The obtained test-retest reliability coefficients were much below the ones derived from the standardization sample and used in establishing the omnibus SEM: VIQ r_{xx} = .48, PIQ r_{xx} = .78, FSIQ r_{xx} = .68. His results are a sobering reminder that the r_{xx} published by the test developers represents the higher limit of the instrument's reliability and is unlikely to be replicated in everyday practice. Consequently, the SEM used in clinical practice is likely to be larger than the one commonly used to compute confidence intervals.

In contrast with the classical measurement models, item response theory (IRT) predicts conditional error variance—specifically, a U-shaped distribution of errors for standard fixed-item tests, with SEMs lowest around the mean ability levels and highest towards the tail of the distribution (Embretson, 1996). This is one version of the conditional

error variance hypothesis. Such a position has face validity because most tests contain few items that can adequately discriminate among individuals with extreme ability levels, which leads to floor and ceiling effects, thus artificially reducing variability. Dimitrov (2002) pointed out that with conditional error variance in a test population, SEM estimates based on different score levels would be more appropriate. He offered a refined formula using integration over the ability continuum, based on the probability density function for the trait distribution, arguing that computations taking into consideration the examinee's ability level allow for more accurate estimates of SEM.

Data are available about the psychometric properties of the WISC-IV at different ability levels. The manual presents statistics for the normative sample in detailed tables, broken down by subtests, age groups, and special categories (Wechsler, 2003). The gifted group (IQ > 130) has a generally lower than average reliability coefficient across subtests, whereas the group with mild retardation has reliability coefficients similar or even higher than the rest of the standardization sample. This finding is also incongruent with the assumption implied in classical test theory that reliability coefficients should be uniform across ability levels, and it provides evidence for another version of the conditional error variance hypothesis, namely, that error is linearly related to ability level. In other words, the normative data of the WISC-IV contradict the assumption of uniform reliability along the ability scale implied by classical test theory and support the hypothesis of conditional error variance.

Given that data from the standardization sample for the WISC-IV imply that reliability varies along the ability scale, it naturally follows that SEM will also change as a

function of IQ. The question of interest is *How does it change—in a linear or curvilinear manner?* There is evidence to support both.

In a study done by Oakland, Lee, and Axelrad (1975), three partially completed WISC protocols of individuals at different ability levels (below average, average, and above average) were mailed to 400 randomly selected psychologists for scoring. A total of 94 participants returned the protocols. The standard deviation of the scores psychologists assigned to each protocol can be interpreted as an empirically derived SEM. The average condition had the smallest overall standard deviation, and the above average condition had the largest one, with the below average condition in between, suggesting that error variance is highest in the right tail of the distribution, reaches its lowest value around the mean, and increases again towards the left tail. This trend was consistent across subtests as well, supporting the U-shaped model of SEM.

Franklin, Stillmann, Burpeau, and Sabers (1982) asked 33 practicing school psychologists and graduate students eligible for state certification as psychometrists to administer the WAIS to four actors who had memorized a script, each of which was designed to yield a different actual IQ. Error rates as measured by the standard deviation of the observed score distribution showed a clear increasing trend from the low to the high ability conditions, with one striking exception: One of the low average conditions (FSIQ = 83) had the same SD as the superior condition (FSIQ = 129). This finding also suggests that error variance increases toward both tails of the distribution.

Other studies found a linear rather than U-shaped relationship between true score and SEM. It logically follows that if there is comparable variability in observed scores along the ability continuum but reliability changes in a linear way, SEM will also change linearly.

Hopwood and Richard (2005) tested scoring accuracy as a function of both stimulus complexity (digitized film clips vs. partially completed record form) and ability level (a FSIQ of 84 vs. 112). No significant stimulus by IQ interaction was found, but participants made significantly more scoring errors in the high IQ condition regardless of the stimulus condition. Although ability was positively related to error, in the absence of a third condition (i.e., a protocol with a FSIQ of 100), a curvilinear trend could not be tested.

It has been shown that contrary to the omnibus error variance assumption, reliability decreases toward the lower tail of the IQ distribution. Using the WISC, Davis (1966) tested 142 children with mental retardation and found that r_{xx} increased gradually as a function of IQ, and, conversely, SEM decreased as IQ increased. This study focused on the lower half of the distribution and concluded that there was an inverse relationship between SEM and observed IQ. However, without an average and a high IQ condition, one cannot determine whether the trend would have reversed.

Webster (1988) performed a test-retest reliability study with the WISC-R on 155 adolescents diagnosed with mild mental retardation (MR, mean IQ = 65.4) and learning disability (LD, mean IQ = 94.2). Even though the correlation between pre- and posttest scores is conceptually independent from internal consistency, SEM, or any other measure of error, it can be used as an indirect index of error variance. Reliability coefficients of the MR group were significantly below those reported in the manual: .57 for FSIQ, .64 for VIQ, and .53 for PIQ. None of the subtest's r_{xx} exceeded .85. The test scores of the LD group proved to be more stable: FSIQ r_{xx} = .99, VIQ r_{xx} = .96, and PIQ r_{xx} = .95. However, Atkinson (1990) pointed out that Webster had used the wrong measure of error variance (SEM instead of SEP, standard error of prediction), thus underestimating the reliability of the WISC-R.

Despite the impressive internal consistency of the WISC-IV (r_{xx} ranging from .95 to .99 according to Wechsler, 2003), its reliability in the field depends on how well individual practitioners adhere to the strict administration and scoring rules. Real-life testing situations are more error prone than the ideal laboratory settings where the test was developed, as the body of empirical research on the WISC attests. No study has replicated the level of reliability reported in the manual. Moreover, as suggested by the studies described above, omnibus indices of reliability may be inaccurate when applied to scores at different ability levels if r_{xx} is a function of ability as predicted by newer statistical models of testing (Dimitrov, 2002, 2003; Embretson, 1996; Lord & Novick, 1968).

Purpose

The present project was inspired by the study done by Hopwood and Richard (2005) and was designed to further investigate the conditional error variance hypothesis. Two important changes were made in the design: adding a third level of the independent variable (the average ability condition) and using more extreme criterion scores (two instead of one standard deviations away from the mean). Hopwood and Richard measured the rate of examiner scoring error using partially completed protocols with fabricated responses that had predetermined Full Scale IQs of 84 and 112. Graduate student participants made significantly more errors in the high IQ condition, (d = .48, partial eta squared = .18). The current study was designed to determine whether scoring errors are a function of examinee ability level. Thus, scoring accuracy was tested at three different ability levels using partially completed protocols that contained only the Vocabulary subtest and were designed to yield scaled scores of 4, 10, and 16 if scored correctly. This subtest was chosen because it has been shown in prior research to yield a disproportionately high number of examiner scoring

errors compared to other subtests (Belk, LoBello, Ray, & Zacar, 2002; Butler, 1954; Kaspar, Throne, & Schulman, 1968; Miller, Chansky, & Gredler, 1970; Slate & Chick, 1989; Slate & Jones, 1990a, 1990b, 1990c; Slate, Jones, Coulter, & Covert, 1992; Slate, Jones, Murray, & Coulter, 1993; Patterson, Slate, Jones, & Steger, 1995; Plumb & Charles, 1955; Vance, Blixt, & Ellis, 1981; Walker, Hunt, & Schwartz, 1965). Because examiners are likely to make the most errors on this subtest, it provided the most powerful test of the conditional error variance hypothesis.

If the rate of scoring errors is significantly different across ability levels within a subtest, then a uniform SEM across ability levels is not justified. In other words, if the SEM is conditional on ability level, confidence intervals should expand or contract as a function of the obtained score's location on the IQ scale.

Hypotheses

It was hypothesized that the SEM is not uniform but is a function of an examinee's true score. Two alternative hypotheses were formulated. According to the linearly increasing measurement error hypothesis, examiners scoring Vocabulary subtests of higher performing individuals will make more item-level errors and be less accurate than examiners scoring subtests of lower performing individuals. Alternatively, the U-shaped (IRT) measurement error hypothesis states that low and high ability levels (4 and 16) will produce similar error rates that are significantly higher than that observed around the medium ability level (10).

Two aspects of error variance were used as dependent variables. Item-level error rate, defined as the ratio of scoring errors to the total number of items scored, was the primary measure of scoring error, as it corrects for the difference in the length of the protocols. A scaled score deviation (criterion score minus obtained score) was used as a

more pragmatic measure of error because this is where scorer judgment starts affecting the final IQ score.

Method

Graduate students in clinical psychology were asked to score WISC-IV Vocabulary subtests for which a fictitious examinee's item responses were provided but had not been scored. Representing three ability levels, the subtests had predetermined true scores of 4, 10, and 16. Participants' completed and returned record forms were then compared to the appropriate criterion scores, and the total number of item-level scoring errors was calculated for each participant and each ability level. The standard deviation associated with each IQ condition can be interpreted as an empirically derived SEM.

The thesis committee approved the proposal of the project on July 15, 2005. The University's Human Subject Research Review Committee approved the project on October 17, 2005 (Approval # 003-06). No data were collected before the day of approval.

Participants

Twenty-eight graduate students (12 men and 16 women) from a Midwestern comprehensive university participated in the study. Participants were recruited informally and through graduate classes. The age range was from 22 to 51 years (M = 28.2, SD = 8.1). Half of the participants were pursuing a Master's degree, and half of them were doctoral students in clinical psychology. All participants completed the same course in administering and scoring the WISC-IV. However, they varied on several variables potentially related to scoring proficiency (length of graduate training [M = 23.3, SD = 15.5 months], number of clinical administrations of the Wechsler scales [M = 13.3, SD = 17.1], and grade received in the assessment course [M = 3.7, SD = .4 on a 4-point scale]), providing a projected ability

range similar to the one found among entry-level psychometricians. Extra credit for participation was provided for some of the participants.

Materials

After having signed the informed consent, each participant scored three protocols that included a fictitious examinee's responses to test items but not his scores. The protocols were developed by the author in consultation with three PhD-level psychologists and a graduate student who had experience with intellectual assessment. Items were revised until a consensus was reached about the correct score. Actual responses of children who had taken the WISC-IV were obtained from archived protocols and were used in developing the protocols.

Three different record forms were used. If scored accurately, each was designed to yield a subtest scaled score of 4, 10, and 16, respectively. A score of 10 is the population average (50^{th} percentile), a score of 4 is two standard deviations below the mean (2^{nd} percentile), and a score of 16 is two standard deviations above the mean (98^{th} percentile). The number of items was increased beyond the typical number of responses expected at the given ability level in the first two conditions (4 & 10) so that participants would not be given cues about the expected score on the basis of the length of the protocol alone. The low ability condition contained 17 items, the average ability condition had 22 items, and the high ability condition had 27 items. Each participant scored three subtests, one at each ability level. Participants were asked to score the items as if they had come from a male who was 11 years and six months old at the time of testing.

Participants were randomly given a sequence of three protocols, one at each of the three ability levels. Thus, each participant's scoring of the protocols can be considered an

independent replication effort of the empirical error rate associated with each ability level. Despite the artificiality of target stimuli (i.e., testing based on fabricated protocols involving no real person), the fact that in this context repeated scoring could be conceived as true parallel measurements enables an important experimental control over extraneous variables typically present with actual examinees. It must be noted, however, that the data points in this analysis were not generated by the same (fictitious) participant but came from repeated scorings of the same protocol by real examiners. Therefore, the measurement analogy used in this study is imperfect, yet perhaps it is the only possible empirical approximation of the distribution of observed scores over repeated attempts to measure a true score given the practical restrictions on intellectual assessment (i.e., learning effects, cost of test administration).

Power Analysis

To detect a medium effect size on error rate with three levels of the IV with a one-way ANOVA while keeping alpha at a .05 level and beta at a .20 level, the sample size required is 52 per condition, or a total of 156 participants (Cohen, 1992). Effect size is defined as the degree to which a phenomenon is present in the population, without implying causality among variables (Cohen, 1988). Medium effect is an informal label introduced by Cohen to describe a relationship between two variables that is detectable by the naked eye of a careful observer. Given that the present study used a repeated measures design, which is more powerful than the independent group contrast, fewer participants were required in order to attain a .80 power. Also, because of the unusually large observed effect, a much smaller sample was needed to detect it.

Response Ambiguity

An ambiguous item is defined as a response that cannot be clearly scored on the basis of the instructions and examples provided in the WISC manual; therefore, the examiner must rely on personal judgment. As ambiguity has been shown to be an important variable in scoring accuracy, it must be controlled for in the protocols (Sattler & Winget, 1970; Sattler, Winget, & Roth, 1969; Slate & Hunnicutt, 1988; Slate & Jones, 1990b; Plumb & Charles, 1955; Walker et al., 1965). To maintain the same absolute level of scoring difficulty among the different IQ conditions, the amount of item ambiguity was held constant across conditions. There were two ambiguous items (one between the score of 0 and 1 and one between the score of 1 and 2) in each condition.

Issues Related to Test Administration

Even though not directly related to scoring accuracy, knowledge of general test administration rules was also assessed. For example, each protocol contained two items that should have been queried on the basis of the scoring manual. Participants were instructed to mark items that they would have queried. Two separate dependent variables emerged from this request: number of correctly identified missing queries (value ranges from 0 to 2) and number of over-queried items (value ranges from 0 to the total number of items in the respective protocol). An instance of over querying was defined either as (a) querying an item that should receive a 2-point score without the extra query or (b) querying an item that already contained a query.

Also, the low ability protocol started with a response that was calibrated as a 1-point answer. On the basis of the standard administration protocol, the reverse rule should have

been applied here. The participant's ability to recognize and mark this rule was also recorded.

Procedures

Partially completed, unscored protocols were distributed to participants for scoring in a quiet office setting at the university. Each participant was provided with three protocols containing only the Vocabulary subtest and a WISC-IV scoring manual. Participants were randomly assigned to one of the possible sequences of the three ability levels (4, 10, and 16). Participants were also asked to assess their overall perceived proficiency in scoring the protocols on a 0-to-100 scale before and after they scored the protocols, as well as to monitor the time taken to finish scoring. This and other demographic variables (sex, age, length and level of training, latency of the Wechsler class and grade obtained, number and latency of clinical WISC-IV/WAIS-III administrations) were used for a correlational analysis to search for scorer variables that systematically covaried with scoring error.

Results

The primary analysis involved repeated measures ANOVAs for each subtest in which the following dependent variables were used: (a) scoring error rate for subtests and (b) the deviation of the observed subtest scaled score from the criterion scaled score. The scoring error rate is the number of errors made by a participant on a subtest divided by the total number of items on the protocol. The observed subtest scaled score is the standardized score of the observed raw score. Mauchly's test of sphericity was performed on all ANOVAs described below. A nonsignificant *W* indicated in each case that the sphericity assumption was met for that analysis.

Item-level scoring errors. The purest measure of scoring error, total item-level error rate is conceptualized as the ratio of incorrectly scored items over the total number of items in a protocol. Its value ranges between 0 (no error) and 1 (every single item was scored wrong). As the proportion of incorrect item-level rater judgment and total number of items scored, the error rate is an unbiased estimator of the error variance, and its values are directly comparable across conditions.

A significant within-subjects effect on item-level scoring errors was observed among the three levels of the IV: $F(2, 54) = 39.09$, $p < .001$, partial eta squared = .59, with an observed power of 1.0. Both the linear [$F(1, 54) = 35.53$, $p < .001$, partial eta squared = .57] and the quadratic [$F(1, 54) = 46.57$, $p < .001$, partial eta squared = .63] effect were significant.

Deviations from the criterion scaled score. We calculated the deviation scores by subtracting the obtained scores (x_i) from the criterion scores (τ). The absolute value of $x_i - \tau$ was treated as an alternative dependent variable. A significant within-subjects effect was

observed in scaled score deviations: $F(2, 54) = 9.87$, $p < .001$, partial eta squared = .28, with an observed power of .98. Both the linear [$F(1, 54) = 6.12$, $p < .05$, partial eta squared = .19, observed power = .66] and the quadratic [$F(1, 54) = 13.31$, $p < .001$, partial eta squared = .34, observed power = .94] effect were significant.

It was hypothesized that the variability in observed scores in the low (4) and high ability (16) conditions would be significantly larger than the variability associated with the average ability (10) condition. The alternative hypothesis also implied that the standard deviation of the observed scaled scores would be greater than the SEM reported in the WISC-IV manual. This hypothesis is generally not supported by the data. As shown in Table 1, despite the strong quadratic effect observed in the sample, the obtained values and the values published in the manual at any given ability level are close to each other. We used Hays's (1973) formula to perform a significance test between the standard deviation associated with each of the three conditions and the SEMs specified by the manual:

$$(N-1)s^2/\delta^2 = \chi^2_{(N-1)}$$

Table 1

Means and Standard Deviations of Scaled Scores as Well as SEMs from the WISC-IV Technical and Interpretive Manual

	Mean	Standard deviation	SEM (11-yo.)	SEM (12-yo.)
Low ability (4)	4.74	1.06	1.08	.90
Average ability (10)	9.89	.96	1.08	.90
High ability (16)	16.21	1.20	1.08	.90

Only one comparison, the variance of the high ability condition compared to the SEM of the 12-year-old group reached statistical significance [χ^2 (27) = 48.00, p < .05]. Although this issue is of theoretical importance and pertinent to the main hypothesis, it has little if any impact on the routine of intelligence testing. In everyday practice, confidence intervals are not computed at subtest level; hence, the true value of the SEM is irrelevant. Replicating this study at the full-scale IQ level, however, could address the question whether a discrepancy exists between the theoretically derived and the empirically produced SEM.

As a measure of variability of observed scores around the criterion score, the sample standard deviation can be considered (by definition) an empirically derived SEM. In this case the protocols are experimental analogues that unlike real examinees, do not change as a result of repeated assessment. An important conceptual difference in the present study is that the replicated measure is not an examinee taking the same test repeatedly but the same test data scored by many different judges.

One-sample *t* tests were performed at each level of the IV to test for the significance of the difference between the obtained mean and the criterion score. Only the low ability level reached statistical significance, $t(27) = 3.69$, $p < .001$, $d = .70$ (medium-large effect size).

Figures 2 and 3 visually represent the results of the present study. Figure 2 graphs the mean item-level error rate across the three ability levels. In essence, it shows a V-shaped distribution of errors, consistent with the conditional error variance hypothesis. On average, in the low ability condition, participants made a scoring error on 20% of the items; in the average ability condition they erred on 6% of the items; and in the high ability condition they made an error on 10% of the items.

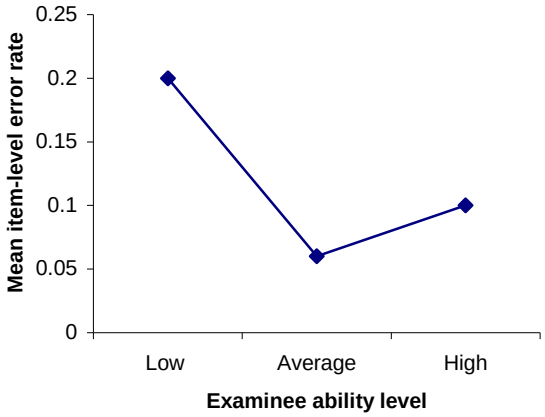

Figure 2. Error distribution expressed in mean item-level error rate.

Figure 3 shows the mean deviation from the criterion score expressed in scaled score units ($M = 10$, $SD = 3$). On average, participants scored the first protocol .71 points higher than the criterion score (4), they scored the second .11 points lower than the criterion score (10), and they scored the third protocol .25 points higher than the pre-established criterion (16).

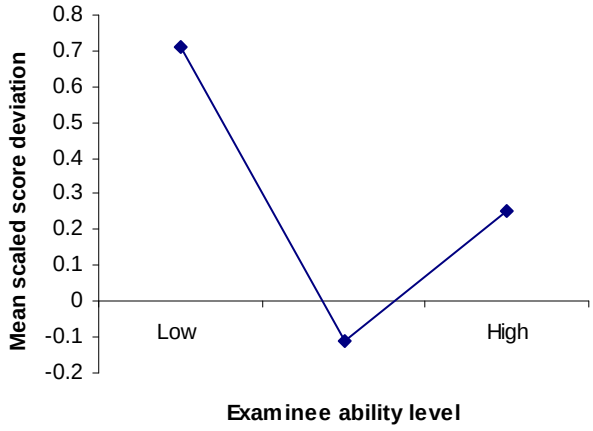

Figure 3. Error distribution expressed in mean scaled score deviation units.

The relationship between scorer variables and error variance. Table 2 shows how some scorer variables relate to the two major measures of error variance. Mean error rate was computed by adding the error rates associated with each level of the independent variable (low, average, and high examinee ability) and dividing it by three. Similarly, the mean scaled score deviation was computed by adding the average deviation from each of the scaled criterion scores and dividing it by three. Although not directly related to the initial hypotheses, this correlational analysis was performed to elucidate the differential contribution of the shown variables to the two measures of error variance.

A number of general observations can be made about the patterns of correlation observed in the data. First, there seems to be no relationship between the two measures of error across the listed variables, which seems puzzling given that theoretically, error rate and scaled score deviation are simply different ways to measure the same thing, that is, error variance (see Table 3). Second, scorer variables tend to correlate more with error rate than with scaled score deviations (Table 2). The Pearson coefficients associated with the former are generally higher than the ones associated with the latter. All four statistically significant correlations were associated with item-level error rate. Third, because of the small sample size and small effect, the matrix is vulnerable to spurious correlations, so even statistically significant values should be interpreted with caution. Even variables that logically should be unrelated to error variance, like subject number, produced correlation coefficients as high as .17. Also, the fact that the two measures of error were sometimes inversely related leads us to believe that at least one of the dependent measures is spuriously related to the scorer variables. Given the above observations, it seems that scaled score deviation is an artificial and unreliable measure of true scoring error.

Table 2

Correlations Among Error Rate, Scaled Score Deviation and Other Measured Variables in the Study

	Mean error rate	Mean scaled score deviation
Age of participant	+.47*	+.06
Latency of Wechsler course	+.45*	-.27
Time taken to complete scoring	+.38*	+.09
Identified missing queries	-.44*	+.20
Subject number	+.17	+.11

* Significant at $p < .05$ level

Table 3 summarizes the correlation between error rate and scaled score deviation at each level of the IV. The last row, labeled Mean, shows the correlation coefficient between the two DVs, averaged for the three levels of the IV. Again, the two measures of error seem to be unrelated. A tentative explanation would be that the conversion of raw scores to scaled scores absorbs some of the item-level errors: Mistakes made in opposite direction cancel each other out without causing any deviation from the true score. For example, if a scorer marks five 1-point answers as 0 and another five 1-point answers as 2, he or she makes 10 mistakes but will obtain the correct scaled score. Conversely, if a scorer gives two points to two 1-point answers and scores the rest of the items correctly, he or she makes only two mistakes, but that will be enough to change (inflate) the scaled score. In other words, 10 item-level errors can go undetected, but 2 may show up at the scaled score level. Therefore, it is not surprising at all that the two measures of error variance are unrelated.

Table 3

The Correlation Between Error Rate and Scaled Score Deviations

	r_{xy}
Low ability (4)	+.12
Average ability (10)	-.14
High ability (16)	-.18
Mean	-.04

Table 4 shows the relationship between querying and error, both at each level of the IV individually and collapsed across them. Only identifying missing queries correlated significantly with mean item-level error rate. Finally, Tables 5 and 6 give a visual summary of the frequency distributions of item-level error rates and scaled score deviations, respectively. They show the frequency of error at each ability level.

Table 4

The Correlation Between Querying and Error

	Condition	Error rate	Scaled score deviation
Identifying missing	Low ability	-.30	+.33
query	Average ability	-.09	+.23
	High ability	-.38*	-.23
	On average	-.44*	+.20

* Significant at $p < .05$ level

Table 5

Frequency Distribution of Item-Level Error Rates at Each of the Ability Levels

	Levels of the IV					
Number of	Low ability		Average ability		High ability	
errors	Frequency	%*	Frequency	%*	Frequency	%*
0	0	0	8	29	0	0
1	3	11	13	46	6	21
2	4	14	2	7	10	36
3	8	29	1	4	4	14
4	8	29	2	7	6	21
5	3	11	1	4	1	4
6	2	7	1	4	1	4

* Percentage may not add up to 100 because of rounding error.

Table 6

Frequency Distribution of Scaled Score Deviations at Each of the Ability Levels

	Levels of the IV					
Deviation	Low ability		Average ability		High ability	
	Frequency	%*	Frequency	%*	Frequency	%*
-3	0	0	1	4	1	4
-2	1	4	1	4	0	0
-1	1	4	5	18	7	25
0	10	36	14	50	8	29
1	10	36	7	25	7	25
2	5	18	0	0	5	18
3	1	4	0	0	0	0

* Percentage may not add up to 100 because of rounding error.

Discussion

The conditional error variance hypothesis was supported by the data. Both measures of error show a V-shaped distribution, and the left side of the V is twice as tall as the right side. In other words, twice as much scoring error was made in the low ability condition than in the high ability condition regardless how error was defined. Even though pairwise comparisons show that all three conditions are significantly different from each other in item-level error rate, it is apparent that the error variance in the low ability condition drives both the quadratic and the linear effect. When contrasted on scaled score deviations, the average and high ability conditions were not different, whereas the low ability condition continued to be different from the other two.

Item-level error rate is a purer measure of scoring error from a theoretical standpoint because it counts each mistake a scorer makes and adjusts it for the total number of items. The amount of deviation from the scaled score, however, is a more pragmatic way to assess scoring accuracy: After all, this is the only way error can influence full-scale IQ and, thus, diagnostic decisions. The two main dependent variables were virtually independent of each other (r_{xy} = -.04). Theoretically, it is possible for a scorer to make a mistake on every single item yet come up with the correct scaled score. For example, if a subtest consists of an even number of 1-point items and someone scores half of them as zero and the other half as two, the person will produce the same scaled score as he or she would by correctly scoring the items. Conversely, given that ambiguous items had two different scores accepted as correct, in this study one could produce a wrong scaled score even without making any item-level errors by giving the two ambiguous items either the higher or the lower score both times.

Scoring errors in opposite direction that cancel each other out have been long reported in the literature. The most important implication of this finding is that random item-level errors, regardless of their frequency, are absorbed by the score conversion procedure, often having no effect on the examinee's final score. On the other hand, even a couple of systematic errors can cause a deviation from the true scaled score. As an artifact of the norm-referenced raw score conversion and statistical weighting, there are two levels of such invisible *error filters* in the Wechsler scales, where item- or subtest-level mistakes are functionally eliminated: first, when transferring raw score totals to subtest scaled scores, as described above, and second, when deriving a full scale IQ from subtest scores. Consequently, random errors have little effect on the final results.

Systematic (i.e., unidirectional) errors, however, influence test scores even at low frequency. The halo effect often reported in scoring accuracy research was also observed in the present study: The only significant deviation from the criterion score (in the low ability condition) was in the positive direction. If this effect is robust and omnipresent in IQ testing, it may contribute to the underdiagnosing of mental retardation. More sophisticated replications of this study are needed to further investigate this issue.

Correlational analyses revealed a few surprising patterns of covariation. Somewhat paradoxically, but consistent with previous studies, older, more experienced participants who were farther along in their graduate training and took more time to complete the scoring tended to make more errors than younger, less experienced students working more quickly. Reliance on the manual negatively correlated with errors, this being the only measured variable that can be easily adjusted to improve scoring performance. The ability to detect missing queries and the reversal rule also had an inverse relationship with error rate, perhaps

because both of those variables are indicative of familiarity with the administration procedure and vigilance while scoring the protocols. Self-reported confidence in scoring did not correlate significantly with any relevant variable. This may be partially due to the restricted range: On a 0-to-100 scale, the lowest score was 70, and the highest was 98.

The present study has several limitations. First, the sample used comes from a single graduate training program; thus, it may not be representative of the general population of graduate students in clinical psychology. Second, and more important, graduate students' scoring abilities may be different in important ways from that of practicing psychometricians. Third, the artificiality of the protocols may compromise the generalizability of the findings: Even though the items themselves come from a pool of actual examinee responses given during real WISC-IV administrations, the subtests were constructed using arbitrary decisions in an attempt to maximize internal validity. Fourth, the subtest with the highest reported error variance was deliberately chosen to magnify patterns of error—ability level covariation. It may be the case that conditional error variance does not operate within other, more objectively scored subtests such as Picture Completion or Block Design.

The present study has two main implications for the practice of intellectual assessment. First, psychometricians may have a tendency to significantly inflate the scores of examinees whose true scores are at the demarcation line between Borderline and Mild Mental Retardation. This can lead to the underdiagnosing of mental retardation. Second, the most important examiner trait that influences test scores is a systematic, one-way bias in scoring. Therefore, assessing and correcting the tendency of practitioners to err in a given direction would be a meaningful part of the screening and training of professionals. Future

studies should be conducted to replicate these findings with different samples and WISC-IV subtests or other assessment instruments.

References

Aiken, L. R. (2000). *Psychological testing and assessment*. Boston, MA: Allyn and Bacon.

Atkinson, L. (1990). Measuring stability and change in WISC-R IQs. *Psychology in the Schools, 27*, 185-186.

Belk, M. S., LoBello, S. G., Ray, G. E., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment, 20*, 290-300.

Butler, A. (1954). Test-retest and split-half reliabilities of the Wechsler-Bellevue scales and subtests with mental defectives. *American Journal of Mental Deficiency, 59*, 80-84.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Davis, L. J., Jr. (1966). The internal consistency of the WISC with the mentally retarded. *American Journal of Mental Deficiency, 70*, 714-716.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27*, 440-458.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*, 783-801.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.

Franklin, M. R., Stillmann, 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.

Friedman, R. (1970). The reliability of the Wechsler Intelligence Scale for Children in a group of mentally retarded children. *Journal of Clinical Psychology, 26*, 181-182.

Gregory, R. J. (1996). *Psychological testing. History, principles and applications*. Boston, MA: Allyn and Bacon.

Hays, W. L. (1973). *Statistics for the social sciences*. New York, NY: Holt, Rinehart and Winston.

Hopwood, C. J., & Richard, D. C. S. (2005). WAIS-III scoring accuracy is a function of scale IC and complexity of examiner tasks. *Assessment, 12*, 445-454.

Kaspar, J. C., Throne, F. M., & Schulman, J. L. (1968). A study of the inter-judge reliability in scoring the responses of aa group of mentally retarded boys to three WISC subscales. *Educational and Psychological Measurement, 28*, 469-477.

Lord, F. M. (1984). Standard errors of measuremeent at different ability levels. *Journal of Educational Measurement, 21*, 239-243.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Miller, C. M., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. *Psychology in the Schools, 7*, 190-193.

Murphy, K. R., & Davidshoffer, C. D. (1998). *Psychological testing. Principles and applications*. Upper Saddle River, NJ: Prentice Hall.

Oakland, T., Lee, S. W., & Axelrad, K. M. (1975). Examiner differences on actual WISC protocols. *Journal of School Psychology, 13*, 227-233.

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

Plumb, G. R., & Charles, D. C. (1955). Scoring difficulty of Wechsler Comprehension responses. *Journal of Educational Psychology, 46*, 179-183.

Sattler, J. M., Winget, B. M., & Roth, R. J. (1969). Scoring difficulty of WAIS and WISC Comprehension, Similarities and Vocabulary responses. *Journal of Clinical Psychology, 25*, 175-177.

Sattler, J.M., & Winget, B. M. (1970). Intelligence testing procedures as affected by expectancy and IQ. *Journal of Clinical Psychology, 26*, 446-448.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools,* 26, 78-84.

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. *The Journal of School 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. (1990c). Student's error in administering the WISC-R: Identifying problem areas. *Measurement and Evaluation in Counseling and Development, 23*, 137-140.

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

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. *Measurement and Evaluation in Counseling and Development, 25*, 156-161.

Vance, H. B., Blixt, S., & Ellis, R. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology, 37*, 397-399.

Walker, R. E., Hunt, W. A., & Schwartz, M. L. (1965). The difficulty of WAIS comprehension scoring. *Journal of Clinical Psychology, 21*, 427-429.

Webster, R. E. (1988). Statistical and individual temporal stability of the WISC-R for cognitively disabled adolescents. *Psychology in the Schools, 25*, 365-372.

Wechsler, D. (2003). *WISC-IV technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.