Psychological Assessment
2014, Vol. 26, No. 1, 207–214

1040-3590/14/$12.00 DOI: 10.1037/a0034832

Whose IQ Is It?—Assessor Bias Variance in High-Stakes Psychological Assessment

Paul A. McDermott
University of Pennsylvania

Marley W. Watkins
Baylor University

Anna M. Rhoad
University of Pennsylvania

Assessor bias variance exists for a psychological measure when some appreciable portion of the score variation that is assumed to reflect examinees' individual differences (i.e., the relevant phenomena in most psychological assessments) instead reflects differences among the examiners who perform the assessment. Ordinary test reliability estimates and standard errors of measurement do not inherently encompass assessor bias variance. This article reports on the application of multilevel linear modeling to examine the presence and extent of assessor bias in the administration of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV) for a sample of 2,783 children evaluated by 448 regional school psychologists for high-stakes special education classification purposes. It was found that nearly all WISC–IV scores conveyed significant and nontrivial amounts of variation that had nothing to do with children's actual individual differences and that the Full Scale IQ and Verbal Comprehension Index scores evidenced quite substantial assessor bias. Implications are explored.

Keywords: measurement bias, assessment, assessor variance, WISC–IV

The Wechsler scales are among the most popular and respected intelligence tests worldwide (Groth-Marnat, 2009). The many scores extracted from a given Wechsler test administration have purported utility for a multitude of applications. For example, as pertains to the contemporary version for school-age children (the Wechsler Intelligence Scale for Children—Fourth Edition [WISC–IV]; Wechsler, 2003), the publisher recommends that resultant scores be used to (a) assess general intellectual functioning; (b) assess performance in each major domain of cognitive ability; (c) discover strengths and weaknesses in each domain of cognitive ability; (d) interpret clinically meaningful score patterns associated with diagnostic groups; (e) interpret the scatter of subtests both diagnostically and prescriptively; (f) suggest classroom modifications and teacher accommodations; (g) analyze score profiles from both an interindividual and intraindividual perspective; and (h) statistically contrast and then interpret differences between pairs of component scores and between individual scores and subsets of multiple scores (Prifitera, Saklofske, & Weiss, 2008; Wechsler, 2003; Weiss, Saklofske, Prifitera, & Holdnack, 2006).

The publisher and other writers offer interpretations for the unique underlying construct meaning (as distinguished from the actual nominal labels) for every WISC–IV composite score, subscore, and many combinations thereof (Flanagan & Kaufman, 2009; Groth-Marnat, 2009; Mascolo, 2009). Moreover, the Wechsler Full Scale IQ (FSIQ) is routinely used to differentially classify mental disability (Bergeron, Floyd, & Shands, 2008; Spruill, Oakland, & Harrison, 2005) and giftedness (McClain & Pfeiffer, 2012), to discover appreciable discrepancies between expected and observed school achievement as related to learning disabilities (Ahearn, 2009; Kozey & Siegel, 2008), and to exclude ability problems as an etiological alternative in the identification of noncognitive disorders (emotional disturbance, communication disabilities, etc.; Kamphaus, Worrell, & Harrison, 2005).

As Kane (2013) has reminded test publishers and users, "the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made" and "more-ambitious claims require more support than less-ambitious claims" (p. 1). At the most fundamental level, the legitimacy of every claim is entirely dependent on the accuracy of test scores in reflecting individual differences. Such accuracy is traditionally assessed through measures of content sampling error (internal consistency estimates) and temporal sampling error (test–retest stability estimates; Allen & Yen, 2001; Wasserman & Bracken, 2013). These estimates are commonplace in test manuals, as incorporated in a standard error of measurement index. It is sometimes assumed that such indexes fully represent the major threats to test score interpretation and use, but they do not (Hanna, Bradley, & Holen, 198l;

This article was published Online First November 4, 2013.

Paul A. McDermott, Graduate School of Education, Quantitative Methods Division, University of Pennsylvania; Marley W. Watkins, Department of Educational Psychology, Baylor University; Anna M. Rhoad, Graduate School of Education, Quantitative Methods Division, University of Pennsylvania.

This research was supported in part by U.S. Department of Education's Institute of Education Sciences Grant R05C050041-05.

Correspondence concerning this article should be addressed to Paul A. McDermott, Graduate School of Education, Quantitative Methods Division, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104-6216. E-mail: drpaul4@verizon.net

Oakland, Lee, & Axelrad, 1975; Thorndike & Thorndike-Christ, 2010; Viswanathan, 2005). Tests administered individually by psychologists or other specialists (in contrast to paper-and-pencil test administrations) are highly vulnerable to error sources beyond content and time sampling. For example, substantial portions of error variance in scores are rooted in the systematic and erratic errors of those who administer and score the tests (Terman, 1918). This is referred to as assessor *bias* (Hoyt & Kerns, 1999; Raudenbush & Sadoff, 2008).

Assessor bias is manifest where, for example, a psychologist will tend to drift from the standardized protocol for test administration (altering or ignoring stopping rules or verbal prompts, mishandling presentation of items and materials, etc.) and erroneously scoring test responses (failure to query ambiguous answers, giving too much or too little credit for performance, erring on time limits, etc.). Sometimes these errors appear sporadically and are limited to a given testing session, whereas other errors will tend to reside more systematically with given psychologists and generalize over a more pervasive mode of unconventional, error-bound, testing practice. Administration and scoring biases, most especially pervasive types, undermine the purpose of testing. Their corrupting effects are exponentially more serious when testing purposes are high stakes, and there is abundant evidence that such biases will operate to distort major score interpretations, to change results of clinical trials, and to alter clinical diagnoses and special education classifications (Allard, Butler, Faust, & Shea, 1995; Allard & Faust, 2000; Franklin, Stillman, Burpeau, & Sabers, 1982; Mrazik, Janzen, Dombrowski, Barford, & Krawchuk, 2012; Schafer, De Santi, & Schneider, 2011).

Recently, Waterman, McDermott, Fantuzzo, and Gadsden (2012) demonstrated research designs to estimate the amount of systematic assessor bias variance carried by cognitive ability scores in early childhood. Well-trained assessors applying individually administered tests were randomly assigned to child examinees, whereafter each assessor tested numerous children. Conventional test-score internal consistency, stability, and generalizability were first supported (McDermott et al., 2009), and thereafter hierarchical linear modeling (HLM) was used to partition score variance into that part conveying children's actual individual differences (the relevant target phenomena in any high-stakes psychological assessment) and that part conveying assessor bias (also known as assessor variance; Waterman et al., 2012). The technique was repeated for other high-stakes assessments in elementary school and on multiple occasions, each application revealing whether assessor variance was relatively trivial or substantial.

This article reports on the application of the Waterman et al. (2012) technique to WISC–IV assessments by regional school psychologists over a period of years. The sample comprises child examinees who were actually undergoing assessment for high-stakes special education classification and related clinical purposes. Whereas the study was designed to investigate the presence and extent of assessor bias variance, it was not designed to pinpoint the exact causes of that bias. Rather, multilevel procedures are used to narrow the scope of probable primary causes and ancillary empirical analyses, and interpretations are used to shed light on the most likely sources of WISC–IV score bias.

Method

Participants

Two large southwestern public school districts were recruited for this study by university research personnel, as regulated by Internal Review Board (IRB) and respective school district confidentiality and procedural policies. School District 1 had an enrollment of 32,500 students and included 31 elementary, eight middle, and six high schools. Ethnic composition for the 2009–2010 academic year was 67.2% Caucasian, 23.8% Hispanic, 4.0% African American, 3.9% Asian, and 1.1% Native American. District 2 served 26,000 students in 2009–2010, with 16 elementary schools, three kindergarten through eighth-grade schools, six middle schools, five high schools, and one alternative school. Caucasian students comprised 83.1% of enrollments, Hispanic 10.5%, Asian 2.9%, African American 1.7%, and other ethnic minorities 1.8%.

Eight trained school psychology doctoral students examined approximately 7,500 student special education files and retrieved pertinent information from all special education files spanning the years 2003–2010, during which psychologists had administered the WISC–IV. Although some special education files contained multiple periodic WISC–IV assessments, only those data pertaining to the first (or only) WISC–IV assessment for a given child were applied for this study; this was used as a measure to enhance comparability of assessment conditions and to avert sources of within-child temporal variance. Information was collected for a total of 2,783 children assessed for the first time via WISC–IV, that information having been provided by 448 psychologists over the study years, with 2,044 assessments collected through District 1 files and 739 District 2 files. The assessments ranged from one to 86 per psychologist ($M = 6.5$, $SD = 13.2$). Characteristics of the examining psychologists were not available through school district files, nor was such information necessary for the statistical separation of WISC–IV score variance attributable to psychologists versus children.

Sample constituency for the 2,783 first-time assessments included 66.0% male children, 78.3% Caucasian, 13.0% Hispanic, 5.4% African American, and 3.3% other less represented ethnic minorities. Ages ranged from 6 to 16 years ($M = 10.3$ years, $SD = 2.5$), where English was the home language for 95.0% of children (Spanish the largest exception at 3.8%) and English was the primary language for 96.7% of children (Spanish the largest exception at 2.3%).

Whereas all children were undergoing special education assessment for the first time using the WISC–IV, 15.7% of those children had undergone prior psychological assessments not involving the WISC–IV (periodic assessments were obligatory under state policy). All assessments were deemed as high stakes, with a primary diagnosis of learning disability rendered for 57.6% of children, emotional disturbance for 11.6%, attention-deficit/hyperactivity disorder for 8.0%, intellectual disability for 2.6%, 12.1% with other diagnoses, and 8.0% receiving no diagnosis. Secondary diagnoses included 10.3% of children with speech impairments and 3.7% with learning disabilities.

Instrumentation

The WISC–IV features 10 core and five supplemental subtests, each with an age-blocked population mean of 10 and standard deviation of 3. The core subtests are used to form four factor indexes, where the Verbal Comprehension Index (VCI) is based on

the Similarities, Vocabulary, and Comprehension subtests; the Perceptual Reasoning Index is based on Block Design, Matrix Reasoning, and Picture Concepts subtests; the Working Memory Index (WMI) on the Digit Span and Letter–Number Sequencing subtests; and the Processing Speed Index (PSI) on the Coding and Symbol Search subtests. The FSIQ is also formed from the 10 core subtests. The factor indexes and FSIQ each retain an age-blocked population mean of 100 and standard deviation of 15. The supplemental subtests were not included in this study because their infrequent application precluded requisite statistical power for multilevel analyses.

Analyses

The eight school psychology doctoral students examined each special education case file and collected WISC–IV scores, assessment date, child demographics, consequent psychological diagnoses, and identity of the examining psychologist. Following IRB and school district requirements, the identity of participating children and psychologists was concealed before data were released to the researchers. Because test protocols were not accessible, nor had standardized observations of test sessions been conducted, it was not possible to determine whether specific scoring errors were present, nor to associate psychologists with specific error types. Rather, test score variability was analyzed via multilevel linear modeling as conducted through SAS PROC MIXED (SAS Institute, 2011).

As a preliminary step to identify the source(s) of appreciable score nesting, a three-level unconditional one-way random effects HLM model was tested for the FSIQ score and each respective factor index and subtest score, where Level 1 modeled score variance between children within psychologists, Level 2 modeled score variance between psychologists within school districts, and Level 3 modeled variance between school districts. This series of analyses sought to determine whether sufficient score variation existed between psychologists and whether this was related to school district affiliation. A second series of multilevel models examined the prospect that because all data had been filtered through a process involving eight different doctoral students, perhaps score variation was affected by the data collection mechanism as distinguished from the psychologists who produced the data. Here, an unconditional cross-classified model was constructed for FSIQ and each factor index and subtest score, with score variance dually nested within doctoral student data collectors and examining psychologists.

Setting aside alternative hypotheses regarding influence of data collectors and school districts, each IQ measure was examined through a two-level unconditional HLM model in which Level 1 represented variation between children within examining psychologists and Level 2 variation between psychologists. The intraclass correlation was derived from the random coefficient for intercepts associated with each model and thereafter converted to a percentage of score variation between psychologists and between children within psychologists.

Because psychologists were not assigned randomly to assess given children (assignment will normally vary as a function of random events, but also as related to which psychologists may more often be affiliated with certain child age cohorts, schools, educational levels, etc.), it seemed reasonable to hypothesize that such nonrandom assignment would potentially result in some systematic characterization of those students assessed by given psychologists. Thus, any systematic patterns of assignments by child demographics could somehow homogenize IQ score variation within psychologists. To ameliorate this potential, each two-level unconditional model was augmented by addition of covariates including child age, sex, ethnicity (minority vs. Caucasian), child primary language (English as a secondary language vs. English as a primary language), and their interactions. The binary covariates were transformed to reflect the percentage of children manifesting a given demographic characteristic as associated with each psychologist, and all the covariates were grand-mean recentered to capture (and control) differences between psychologists (Hofmann & Gavim, 1998). Covariates were added systematically to the model for each IQ score so as to minimize Akaike's information criterion (AIC; as recommended by Burnham & Anderson, 2004), and only statistically significant effects were permitted to remain in final models (although nonsignificant main effects were permitted to remain in the presence of their significant interactions). Whereas final models were tested under restricted maximum-likelihood estimation, and are so reported, the overall statistical consequence of the covariate augmentation for each model was tested through likelihood ratio deviance tests contrasting each respective unconditional and final conditional model under full maximum-likelihood estimation (per Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). In essence, the conditional models operated to correct estimates of between-psychologists variance (obtained through the initial unconditional models) for the prospect that some of that variance was influenced by the nonrandom assignment of psychologists to children.

Results

A preliminary unconditional HLM model was applied for FSIQ and each respective factor index and subtest score, where children were nested within psychologists and psychologists within school districts. The coefficient for random intercepts of children nested within psychologists was statistically significant for almost all models, but the coefficient for psychologists nested within districts was nonsignificant for every model. Similarly, a preliminary multilevel model for each IQ score measured cross-classified children nested within data collectors as well as psychologists. No model produced a statistically significant effect for collectors, whereas most models evinced a significant effect for psychologists. Therefore, school district and data collection effects were deemed inconsequential, and subsequent HLM models tested a random intercept for nesting within psychologists only.

For each IQ score, two-level, unconditional and conditional HLM models were constructed, initially testing the presence of psychologist assessor variance and thereafter controlling for differences in child age, sex, ethnicity, language status, and their interactions. Table 1 reports the statistical significance of the assessor variance effect for each IQ score and the estimated percentage of variance associated exclusively with psychologists versus children's individual differences. The last column indicates the statistical significance of the improvement of the conditional model (controlling for child demographics) over the unconditional model for each IQ measure. Where these values are nonsignificant, understanding is enhanced by interpreting percentages associated

Table 1
Percentages of Score Variance Associated With Examiner Psychologists Versus Children's Individual Differences on the Wechsler Intelligence Scale for Children—Fourth Edition

IQ score	*N*	Unconditional models[a]		Conditional models[b]		Difference between unconditional and conditional models (*p*)[c]
		% variance between psychologists	% variance between children	% variance between psychologists	% variance between children	
Full Scale IQ	2,722	16.2***	83.8	**12.5*** **	**87.5**	.0049
Verbal Comprehension Index	2,783	14.0***	86.0	**10.0*** **	**90.0**	<.0001
Similarities	2,551	10.6***	89.4	**7.4*** **	**92.6**	.0069
Vocabulary	2,538	**14.3*** **	**85.7**	10.4***	89.6	*ns*
Comprehension	2,524	**10.7*** **	**87.3**	9.9***	90.1	*ns*
Perceptual Reasoning Index	2,783	**7.1****	**92.9**	5.7**	94.3	*ns*
Block Design	2,544	**5.3****	**94.7**	3.8*	96.2	*ns*
Matrix Reasoning	2,520	**2.8**	**97.2**	2.4	97.6	*ns*
Picture Concepts	2,540	**5.4***	**94.6**	4.9*	95.1	*ns*
Working Memory Index	2,782	9.8***	90.2	**8.3*** **	**91.7**	.002
Digit Span	2,548	**7.8*** **	**92.2**	7.5***	92.5	*ns*
Letter–Number Sequencing	2,486	**5.2***	**94.8**	4.2*	95.8	*ns*
Processing Speed Index	2,778	12.6***	87.4	**7.6*** **	**92.4**	<.0001
Coding	2,528	9.2***	90.8	**4.4***	**95.6**	<.0001
Symbol Search	2,521	**12.7*** **	**87.3**	9.9***	90.1	*ns*

[a] Entries for percentage of variance between psychologists equal ICC × 100 as derived in hierarchical linear modeling. Percentages of variance between children equal (1 − ICC) × 100. Boldface entries are regarded optimal for interpretation purposes (in contrast to entries under the alternative conditional model, which do not represent significant improvement). Model specification is $Y_{ij} = \gamma_{oo} + \mu_{oj} + r_{ij}$, where *i* indexes children within psychologists and *j* indexes psychologists. Significance tests indicate statistical significance of the random coefficient for psychologists, where *p* values > .01 are considered nonsignificant. ICC = interclass correlation coefficient. [b] Entries for percentage of variance between psychologists equal residual ICC × 100 as derived in hierarchical linear modeling, incorporating statistically significant fixed effects for child age, sex, ethnicity, language status, and their interactions. Percentages of variance between children equal (1 −residual ICC) × 100. Boldface entries are regarded optimal for interpretation purposes (in contrast to entries under the alternative unconditional model). Model specification is $Y_{ij} = \gamma_{oo} + \gamma_{o1}MeanAge_j + \gamma_{o2}MeanPercentMale_j + \gamma_{o3}MeanPercentMinority_j + \gamma_{o4}MeanPercentESL_j + \gamma_{o5}(MeanAge_j)(MeanPercentMale_j) + \ldots + r_{ij}$, where *i* indexes children within psychologists, *j* indexes psychologists, and nonsignificant terms are dropped from models. Significance tests indicate statistical significance of the residualized random coefficient for psychologists, where *p* values > .01 are considered nonsignificant. [c] Values are based on tests of the deviance between −2 log likelihood estimates for respective unconditional and conditional models under full maximum-likelihood estimation. *p*s > .01 are considered nonsignificant (*ns*).
* $p < .01$. ** $p < .001$. *** $p < .0001$.

with the unconditional model, and where values are significant, interpretation is enhanced by percentages associated with the conditional model. Following this logic, percentages preferred for interpretation are boldfaced.

The conditional models (which control for child demographics) make a difference for FSIQ, VCI (especially its Similarities subtest), WMI, and PSI (especially its Coding subtest) scores. This suggests at least that the nonrandom assignment of school psychologists to children may result in imbalanced distributions of children by their age, sex, ethnicity, and language status. This in itself is not problematic and likely reflects the realities of requisite quasi-systematic case assignment within school districts. Thus, psychologists will be assigned partly on the basis of their familiarity with given schools, levels of expertise with age cohorts, travel convenience, and school district administrative divisions—all factors that would tend to militate demographic differences across case loads. The conditional models accommodate for that prospect. At the same time, it should be recognized that the control mechanisms in the conditional models are also probably overly conservative because they will inadvertently control for assessor bias arising as a function of children's demographic characteristics (race, sex, etc.) unrelated to case assignment methods.

Considering the major focus of the study (identification of that portion of IQ score variation that without mitigation has nothing to do with children's actual individual differences), the FSIQ and all four factor index scores convey significant and nontrivial (viz. ≥5%) assessor bias. More troubling, bias for FSIQ (12.5%) and VCI (10.0%) is substantial (≥10%). Within VCI, the Vocabulary subtest (14.3% bias variance) and Comprehension subtest (10.7% bias variance) are the primary culprits, each conveying substantial bias. Further problematic, under PSI, the Symbol Search subtest is laden with substantial bias variance (12.7%).

On the positive side, the Matrix Reasoning subtest involves no statistically significant bias (2.8%). Additionally, the Coding subtest, although retaining a statistically significant amount of assessor variance, essentially yields a trivial (<5%) amount of such variance (4.4%). (Note that the <5% criterion for deeming hierarchical cluster variance as practically inconsequential comports with the convention recommended by Snijders & Baker, 1999, and Waterman et al., 2012.)

Discussion

The degree of assessor bias variance conveyed by FSIQ and VCI scores effectively vitiates the usefulness of those measures for differential diagnosis and classification, particularly in the vicinity of the critical cut points ordinarily applied for decision making. That is, to the extent that decisions on mental deficiency and intellectual giftedness will depend on discovery of FSIQs < 70 or ≥ 130, respectively, or that ability-achievement discrepancies (whether based on regression modeling or not) will depend on accurate measurement of the FSIQ, those decisions cannot be

rendered with reasonable confidence because the IQ measures reflect substantial proportions of score variation emblematic of differences among examining psychologists rather than among children. The folly of basing decisions in part or in whole on such IQ measures is accentuated where the evidence (for intellectual disability, etc.) is anything but incontrovertible because the FSIQ score is markedly above or below the cut point or the ability-achievement discrepancy is so immense as to leave virtually no doubt that real and substantial disparity exists (see also Franklin et al., 1982; Gresham, 2009; Lee, Reynolds, & Willson, 2003; Mrazik et al., 2012; Reynolds & Milam, 2012, on the matter of high-stakes decisions following IQ test administration and scoring errors).

This study is limited by virtue of its dependence on a regional rather than a more representative national sample. Indeed, future research should explore the broader generalization of assessor bias effects. From one perspective, it would seem ideal if psychologists could be randomly assigned to children because that process would equitably disperse the myriad elements of variance that can neither be known nor controlled. From another perspective, random assignment is probably infeasible because, to the extent that participant children and their families and schools are expecting psychological services from those practitioners who have the best relationships with given schools or school personnel or expertise with certain levels of child development, the reactivity associated with random assignment for high-stakes assessments could do harm or be perceived as doing harm.

Unfortunately, test protocols were inaccessible, and there were no standardized test session observations. Thus, it was not possible to associate specific errors with specific psychologists. Likewise, there was no information about the psychologists' demographic characteristics, nor the relationship between psychologists and children. However, the magnitude of bias effects found in this study makes it clear that future research should identify the causes of assessor variance and, if feasible, design interventions to reduce bias in children's test scores.

It may be suggested with respect to the current study that the statistical models to control for nonrandom assignment using child demographics as covariates might further have been augmented through covariates controlling for the diagnoses rendered for children. The central hypothesis would be that nonrandom assignment may result in systematic patterns of diagnosis between psychologists, and those differences may be applied to explain why psychologists differ in IQ score generation. We resisted this notion because it essentially reverses the natural causal order whereby IQ scores are first generated and are expected to influence diagnoses, not the reverse where summary diagnoses are expected to influence IQ scores. While pondering the propriety of the hypothesis, we explored the utility of diagnoses to control for differences among psychologists and observed that the use of such information did little to diminish the assessor bias effects discovered without that information.

Given the serious implications of assessor bias for high-stakes psychological assessment, it would be somewhat comforting to believe that the problem might be mitigated through more competent and continued training of practicing psychologists. Thus, psychologists would be trained to a high criterion of accuracy and periodically requalified or refreshed as are some other specialists. This prospect follows the *practice makes perfect* paradigm. For the McDermott et al. (2009) studies, it was apparent that refresher training for those who individually administered cognitive tests tended to motivate trivial levels of assessor variance. But in those studies, test items and procedures that gave common cause for generating errors were systematically altered or eliminated from the tests. This has not been the practice for the WISC–IV, nor for similar Wechsler tests, and, regrettably, there exists rather compelling evidence that additional training does not appreciably mitigate many administration and scoring errors (Kuentzel, Hetterscheidt, & Barnett, 2011; Legris, 2004; Loe, Kadulbek, & Marks, 2007; Moon, Blakey, Gorsuch, & Fantuzzo, 1991; Mrazik et al., 2012; Patterson, Slate, Jones, & Steger, 1995; Slate & Jones, 1990).

Related evidence suggests that certain Wechsler tasks, especially those requiring a uniform standard for the administration and scoring on the various verbal tests (Similarities, Vocabulary, Comprehension) and the consequent formation of dependent measures such as the VCI and FSIQ are simply too complex and make implausible any minimally acceptable accuracy. For example, verbal subtests have been found to be especially susceptible to errors for both graduate students and clinicians (Babad, Mann, & Mar-Hayim, 1975; Beasley, Lobasher, Henley, & Smith, 1988; Belk, LoBello, Ray, & Zachar, 2002; Bradley, Hanna, & Lucas, 1980; Erdodi, Richard, & Hopwood, 2009; Franklin et al., 1982; LoBello & Holley, 1999; Loe et al., 2007; Mrazik et al., 2012; Oakland et al., 1975; Plumb, 1955; Ryan & Schnakenberg-Ott, 2003; Sattler, Winget, & Roth, 1969; Slate & Chick, 1989; Slate, Jones, Murray, & Coulter, 1993).

Moreover, it should be recognized that compromised administration and scoring is not unique to cognitive tests, nor restricted to verbal components of the WISC–IV, but is rather more endemic to psychological assessment in general, reaching back nearly a century and affecting a broad collection of measuring devices (Allard & Faust, 2000; Charter, Walden, & Padilla, 2000; Edwards & Rottman, 2011; Goddard, Simons, Patton, & Sullivan, 2004; Kozora, Kongs, Hampton, & Zhang, 2008; Matthey, Lee, Črnčec, & Trapolini, 2013; Ramos, Alfonso, & Schermerhorn, 2009; Schafer et al., 2011; Simons, Goddard, & Patton, 2002; Sullivan, 2000; Terman, 1918). Evidently, scoring and administration errors, use of the wrong tables, clerical errors, judgment errors, and the like, are pervasive in psychological assessment. Deviations from standardized administration and scoring procedures, which can introduce serious error, may also be ubiquitous (Wolfe-Christensen & Callahan, 2008).

Additionally, characteristics of the examiner, examinee, or examiner–examinee relationship may impact test scores (Glutting, Oakland, & Konold, 1994; see also Sattler, 2008, pp. 41–44). First, psychological examiners are vulnerable to the same cognitive limitations and biases as other humans (Garb, 2010; Ruscio, 2007). Second, test scores may be influenced by the examinee's familiarity with the examiner. Children with language handicaps, learning disabilities, and autism as well as children from low-socioeconomic and minority households have been found to achieve lower scores on demanding cognitive tests if tested by unfamiliar examiners (Cohen's d = .47–.73; Fuchs & Fuchs, 1986, 1989; Fuchs, Fuchs, Garwick, & Featherstone, 1983; Fuchs, Fuchs, & Power, 1987; Fuchs, Fuchs, Power, & Dailey, 1985; Szarko, Brown, & Watkins, 2013). Finally, examinee motivation may impact test scores (Etherton & Axelrod, 2013; Phay, 1990).

Cognitive tests are assumed to be measures of maximal performance, and examinees should be optimally motivated to perform well (Ackerman & Heggestad, 1997). However, a recent meta-analysis revealed that material incentives increased IQ scores by 0.64 standard deviations, suggesting that motivation and effort may not be uniformly high in all testing situations (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). In short, "there are innumerable sources of error in giving and scoring mental tests of whatever kind" (Terman, 1918, p. 33).

Conclusion

Ordinary reliability estimates and consequent standard errors of measurement do not inherently account for assessor bias variance because those statistics only reflect sampling error or temporal instability (Bradley et al., 1980). The many interpretation schemes recommended to compare and contrast factor indexes and their component subtests generally rely exclusively on such error estimates (Glass, Ryan, & Charter, 2010; Glass, Ryan, Charter, & Bartels, 2009; Hanna et al., 1981). Consequently, the nontrivial and substantial amounts of assessor bias that plague almost all factor index and subtest scores effectively diminishes the legitimacy of analyses of score patterns, profiles, or assessments of relative intellective strengths and weaknesses. There is simply too much score variation that has nothing to do with actual differences between (or within) children and too much variation that is fundamentally errant to differential description, let alone to differential diagnosis or classification.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219–245. doi:10.1037/0033-2909.121.2.219

Ahearn, E. M. (2009). State eligibility requirements for specific learning disabilities. *Communication Disorders Quarterly, 30,* 120–128. doi:10.1177/1525740108325221

Allard, G., Butler, J., Faust, D., & Shea, M. T. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaire–Revised (PDQ-R). *Professional Psychology: Research and Practice, 26,* 304–308. doi:10.1037/0735-7028.26.3.304

Allard, G., & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7,* 119–129. doi:10.1177/107319110000700203

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

Babad, E. Y., Mann, M., & Mar-Hayim, M. (1975). Bias in scoring the WISC subtests. *Journal of Consulting and Clinical Psychology, 43,* 268. doi:10.1037/h0076368

Beasley, M. G., Lobasher, M., Henley, S., & Smith, I. (1988). Errors in computation of WISC and WISC-R intelligence quotients from raw scores. *Journal of Child Psychology and Psychiatry, 29,* 101–104. doi:10.1111/j.1469-7610.1988.tb00693.x

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. doi:10.1177/073428290202000305

Bergeron, R., Floyd, R. G., & Shands, E. I. (2008). States' eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities, 43,* 123–131.

Bradley, F. O., Hanna, G. S., & Lucas, B. A. (1980). The reliability of scoring the WISC-R. *Journal of Consulting and Clinical Psychology, 48,* 530–531. doi:10.1037/0022-006X.48.4.530

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33,* 261–304. doi:10.1177/0049124104268644

Charter, R. A., Walden, D. K., & Padilla, S. P. (2000). Too many simple clerical scoring errors: The Rey Figure as an example. *Journal of Clinical Psychology, 56,* 571–574. doi:10.1002/(SICI)1097-4679(200004)56:4<571::AID-JCLP10>3.0.CO;2-6

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108,* 7716–7720. doi:10.1073/pnas.1018601108

Edwards, O. W., & Rottman, A. (2011). Empirical analysis of the relationship between student examiners' learning with deliberate test practice and examinees' intelligence test performance. *Journal of Instructional Psychology, 38,* 157–163.

Erdodi, L. A., Richard, D. C. S., & Hopwood, C. (2009). Importance of relying on the manual: Scoring error variance in the WISC-IV vocabulary subtest. *Journal of Psychoeducational Assessment, 27,* 374–385. doi:10.1177/0734282909332913

Etherton, J. L., & Axelxrod, B. N. (2013). Do administration instructions alter optimal neuropsychological test performance? Data from healthy volunteers. *Applied Neuropsychology: Adult, 20,* 15–19. doi:10.1080/09084282.2012.670152

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Franklin, M. R., Jr., Stillman, P. L., Burpeau, M. Y., & Sabers, D. L. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools, 19,* 563–569. doi:10.1002/1520-6807(198210)19:4<563::AID-PITS2310190427>3.0.CO;2-Q

Fuchs, D., & Fuchs, L. S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research, 56,* 243–262. doi:10.3102/00346543056002243

Fuchs, D., & Fuchs, L. S. (1989). Effects of examiner familiarity on Black, Caucasian, and Hispanic children: A meta-analysis. *Exceptional Children, 55,* 303–308.

Fuchs, D., Fuchs, L. S., Garwick, D. R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. *Journal of Psychology: Interdisciplinary and Applied, 114,* 37–46. doi:10.1080/00223980.1983.9915393

Fuchs, D., Fuchs, L. S., & Power, M. H. (1987). Examiner familiarity on LD and MR students' language performance. *Remedial and Special Education, 8,* 47–52. doi:10.1177/074193258700800407

Fuchs, D., Fuchs, L. S., Power, M. H., & Dailey, A. M. (1985). Bias in the assessment of handicapped children. *American Educational Research Journal, 22,* 185–198. doi:10.3102/00028312022002185

Garb, H. N. (2010). The social psychology of clinical judgment. In J. E. Maddux & J. P. Tangney (Eds.), *Social psychological foundations of clinical practice* (pp. 297–311). New York, NY: Guilford Press.

Glass, L. A., Ryan, J. J., & Charter, R. A. (2010). Discrepancy score reliabilities in the WAIS-IV standardization sample. *Journal of Psychoeducational Assessment, 28,* 201–208. doi:10.1177/0734282909346710

Glass, L. A., Ryan, J. J., & Charter, R. A., & Bartels, J. M. (2009). Discrepancy score reliabilities in the WISC-IV standardization sample. *Journal of Psychoeducational Assessment, 27,* 138–144. doi:10.1177/0734282908325158

Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to Assessment of Test-Session Behavior for the WISC-III and WIAT: Possible ethnicity, gender, and SES effects. *Journal of School Psychology, 32,* 355–369. doi:10.1016/0022-4405(94)90033-7

Goddard, R., Simons, R., Patton, W., & Sullivan, K. (2004). Psychologist hand-scoring error rates on the Rothwell-Miller Interest Blank: A comparison of three job allocation systems. *Australian Journal of Psychology, 56,* 25–32. doi:10.1080/00049530410001688100

Gresham, F. M. (2009). Interpretation of intelligence test scores in *Atkins* cases: Conceptual and psychometric issues. *Applied Neuropsychology, 16,* 91–97. doi:10.1080/09084280902864329

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology, 19,* 370–376. doi:10.1016/0022-4405(81)90031-5

Hofmann, D. A., & Gavim, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24,* 623–641.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403–424. doi:10.1037/1082-989X.4.4.403

Kamphaus, R., Worrell, F. C., & Harrison, P. (2005). Principles for evaluation and eligibility determination for specific learning disabilities: A report of the ad hoc committee of division 16. *The School Psychologist, 59,* 157–159.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. doi:10.1111/jedm.12000

Kozey, M., & Siegel, L. S. (2008). Definitions of learning disabilities in Canadian provinces and territories. *Canadian Psychology, 49,* 162–171. doi:10.1037/0708-5591.49.2.162

Kozora, E., Kongs, S., Hampton, M., & Zhang, L. (2008). Effects of examiner error on neuropsychological test results in a multi-site study. *Clinical Neuropsychologist, 22,* 977–988. doi:10.1080/13854040701679025

Kuentzel, J. G., Hetterscheidt, L. A., & Barnett, D. (2011). Testing intelligently includes double-checking Wechsler IQ scores. *Journal of Psychoeducational Assessment, 29,* 39–46. doi:10.1177/0734282910362048

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology, 3,* 55–81. doi:10.1300/J151v03n03_04

Legris, Y. (2004). Development of a new teaching method to reduce scoring errors on the WISC-III and WAIS-III. *Dissertation Abstracts International: Section A. Humanities and Social Sciences, 65*(1-A), 68.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS systems for mixed models* (2nd ed.). Cary, NC: SAS Institute.

LoBello, S. G., & Holley, G. (1999). WPPSI-R administration, clerical, and scoring errors by student examiners. *Journal of Psychoeducational Assessment, 17,* 15–23. doi:10.1177/073428299901700102

Loe, S. A., Kadulbek, R. M., & Marks, W. J. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. *Journal of Psychoeducational Assessment, 25,* 237–247. doi:10.1177/0734282906296505

Mascolo, J. T. (2009). Appendix 1: Linking WISC-IV assessment results to educational strategies and instructional supports. In D. Flanagan & A. Kaufman (Eds.), *Essentials of WISC-IV assessment* (2nd ed., pp. 1–17). Hoboken, NJ: Wiley.

Matthey, S., Lee, C., Črnčec, R., & Trapolini, T. (2013). Errors in scoring the Edinburgh Postnatal Depression scale. *Archives of Women's Mental Health, 16,* 117–122. doi:10.1007/s00737-012-0324-9

McClain, M.-C., & Pfeiffer, S. (2012). Identification of gifted students in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology, 28,* 59–88. doi:10.1080/15377903.2012.643757

McDermott, P. A., Fantuzzo, J. F., Waterman, C., Angelo, L. A., Warley, H. W., Gadsden, V. L., & Zhang, X. (2009). Measuring preschool cognitive growth while it's still happening: The Learning Express. *Journal of School Psychology, 47,* 337–366. doi:10.1016/j.jsp.2009.07.002

Moon, G. W., Blakey, W. A., Gorsuch, R. L., & Fantuzzo, J. W. (1991). Frequent WAIS-R administration errors: An ignored source of inaccurate measurement. *Professional Psychology: Research and Practice, 22,* 256–258. doi:10.1037/0735-7028.22.3.256

Mrazik, M., Janzen, T. M., Dombrowski, S. C., Barford, S. W., & Krawchuk, L. L. (2012). Administration and scoring errors of graduate students learning the WISC-IV: Issues and controversies. *Canadian Journal of School Psychology, 27,* 279–290. doi:10.1177/0829573512454106

Oakland, T., Lee, S. W., & Axelrad, K. M. (1975). Examiner differences on actual WISC protocols. *Journal of School Psychology, 13,* 227–233. doi:10.1016/0022-4405(75)90005-9

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,* 32–37. doi:10.1177/0013164495055001003

Phay, A. J. (1990). Shipley Institute of Living Scale: Part 1–Moderator variables. *Medical Psychotherapy, 3,* 1–15.

Plumb, G. R. (1955). Scoring difficulty of Wechsler comprehension responses. *Journal of Educational Psychology, 46,* 179–183. doi:10.1037/h0046974

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (2008). *WISC-IV clinical assessment and intervention* (2nd ed.). San Diego, CA: Academic Press.

Ramos, E., Alfonso, V. C., & Schermerhorn, S. M. (2009). Graduate students' administration and scoring errors on the Woodcock-Johnson III tests of cognitive abilities. *Psychology in the Schools, 46,* 650–657. doi:10.1002/pits.20405

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1,* 138–154. doi:10.1080/19345740801982104

Reynolds, C. R., & Milam, D. A. (2012). Challenging intellectual testing results. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony* (6th ed., pp. 311–334). New York, NY: Oxford University Press.

Ruscio, J. (2007). The clinician as subject. In S. O. Lilienfeld & W. T. O'Donohue (Eds.), *The great ideas of clinical science: 17 principles that every mental health professional should understand* (pp. 29–47). New York, NY: Routledge.

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). *Assessment, 10,* 151–159. doi:10.1177/1073191103010002006

SAS Institute. (2011). SAS version 9.3 for Windows [Computer program]. Cary, NC: Author.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

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. doi:10.1002/1097-4679(196904)25:2<175::AID-JCLP2270250217>3.0.CO;2-0

Schafer, K., De Santi, S., & Schneider, L. S. (2011). Errors in ADAS-cog administration and scoring may undermine clinical trials results. *Current Alzheimer Research, 8,* 373–376. doi:10.2174/156720511795745357

Simons, R., Goddard, R., & Patton, W. (2002). Hand-scoring error rates in psychological testing. *Assessment, 9,* 292–300. doi:10.1177/1073191102009003008

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools, 26,* 78–84. doi:10.1002/1520-6807(198901)26:1<78::AID-PITS2310260111>3.0.CO;2-5

Slate, J. R., & Jones, C. H. (1990). Identifying students' errors in administering the WAIS-R. *Psychology in the Schools, 27,* 83–87. doi:10.1002/1520-6807(199001)27:1<83::AID-PITS2310270112>3.0.CO;2-1

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.

Snijders, T., & Baker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Spruill, J., Oakland, T., & Harrison, P. (2005). Assessment of mental retardation. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist–practitioner perspectives* (pp. 299–331). San Diego, CA: Elsevier. doi:10.1016/B978-012564931-5/50010-4

Sullivan, K. (2000). Examiners' errors on the Wechsler Memory Scale-Revised. *Psychological Reports, 87,* 234–240.

Szarko, J. E., Brown, A. J., & Watkins, M. W. (2013). Examiner familiarity effects for children with autism spectrum disorders. *Journal of Applied School Psychology, 29,* 37–51. doi:10.1080/15377903.2013.751475

Terman, L. M. (1918). Errors in scoring Binet tests. *Psychological Clinic, 12,* 33–39.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson.

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, 2nd ed., pp. 50–80). Hoboken, NJ: Wiley.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway? *Early Childhood Research Quarterly, 27,* 46–54. doi:10.1016/j.ecresq.2011.06.003

Wechsler, D. (2003). Interpretive considerations. *In WISC-IV technical and interpretive manual* (pp. 99–108). San Antonio, TX: Psychological Corporation.

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV advanced clinical interpretation*. San Diego, CA: Academic Press.

Wolfe-Christensen, C., & Callahan, J. L. (2008). Current state of standardization adherence: A reflection of competency in psychological assessment. *Training and Education in Professional Psychology, 2,* 111–116. doi:10.1037/1931-3918.2.2.111

Received July 9, 2013
Revision received September 11, 2013
Accepted September 13, 2013 ■