Educational and Psychological Measurement
70(4) 613–627

DOI: 10.1177/0013164409355686
http://epm.sagepub.com
SAGE

Differential Relationships Between WISC-IV and WIAT-II Scales: An Evaluation of Potentially Moderating Child Demographics

Timothy R. Konold[1] and Gary L. Canivez[2]

Abstract

Considerable debate exists regarding the accuracy of intelligence tests with members of different groups. This study investigated differential predictive validity of the *Wechsler Intelligence Scale for Children–Fourth Edition*. Participants from the WISC-IV–WIAT-II standardization linking sample ($N = 550$) ranged in age from 6 through 16 years ($M = 11.6$, $SD = 3.2$) and varied by the demographic variables of gender, race/ethnicity (Caucasian, African American, and Hispanic), and parent education level (8-11, 12, 13-15, and 16 years). Full Scale IQ and General Ability Index scores from the WISC-IV were used to predict scores on Mathematics, Oral Language, Reading, Written Language, and the total composite on the *Wechsler Individual Achievement Test–Second Edition*. Differences in prediction were evaluated between demographic subgroups via Potthoff's technique. Of the 30 simultaneous tests, 25 revealed no statistically significant between group differences. The remaining statistically significant differences were found to have little practical or clinical influence when effect size estimates were considered. Results are discussed in the context of other ability measures that were previously investigated for differential validity as well as educational implications for clinicians.

Keywords

bias, WISC-IV, differential prediction, assessment

[1]University of Virginia, Charlottesville, USA
[2]Eastern Illinois University, Charleston, USA

Corresponding Author:
Timothy R. Konold, University of Virginia, 405 Emmet Street South, PO Box 400277, Charlottesville, VA 22904-4277
Email: konold@virginia.edu

Perhaps no other construct in psychology or education has elicited as much debate as the question of what constitutes intelligence, how one might go about measuring it, and whether the resulting scores are equitable across different groups. The first attempt at measuring human intelligence can be traced back to the early 1800s and the work of Sir Francis Galton. Galton's early attempts at measuring intelligence were met with criticism and largely failed to stand the test of time. This was most likely the result of a failure to formally understand and define the construct of intelligence that was the focus of measurement. Modern theories of intelligence are rooted in the theoretical work of Alfred Binet, Victor Henri, and Theodore Simon that took hold in the mid- to late-1800s. Binet's early theories were operationalized in the Binet–Simon Intelligence Scale (1905)—an instrument that was largely successful in identifying children with mental retardation. Soon to follow were the group administered Army Alpha and Army Beta scales. The eventual declassification of these tests led to a proliferation of commercially available intelligence tests through the mid 1900s, including the first Scholastic Aptitude Test (1926).

In 1949, David Wechsler introduced the first version of the Wechsler Intelligence Scale for Children (WISC), following the original publication of the Wechsler–Bellevue Intelligence Scale for adults (Wechsler, 1939). The success of this instrument spurred numerous revisions including the WISC–Revised (WISC-R; Wechsler, 1974), the WISC–Third Edition (WISC-III; Wechsler, 1991), and most recently, the WISC–Fourth Edition (WISC-IV; Wechsler, 2003). See Wasserman and Tulsky (2005) for a more detailed historical account of the origins of intellectual assessment.

Surveys of contemporary usage reveal that intelligence tests are among the most popular measures administered by psychologists (Stinnett, Havey, & Oehler-Stinnett, 1994; Wilson & Reschly, 1996), and that the Wechsler scales figure most prominently into this arsenal (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Alfonso & Pratt, 1997; Belter, & Piotrowski, 2001; Hutton, Dubes, & Muir, 1992; Kaufman & Lichtenberger, 2000; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Such wide spread utilization exists because these tests have an impressive record of psychometric quality, including their applied utility in projecting student achievement (Bracken & Walker, 1997; Brody, 2002; Brown, Reynolds, & Whitaker, 1999; Flanagan, Andrews & Genshaft, 1997; Naglieri Bornstein, 2003). Intelligence quotient (IQ) tests have a rich history of accounting for meaningful levels of achievement variance (Brody, 2002; Naglieri & Bornstein, 2003), with average IQ–achievement correlations near .55 across age groups (Board of Scientific Affairs of the American Psychological Association, 1996; Brody, 2002). In fact, it is often said that the most important application of intelligence tests is their ability to forecast student achievement (Brown et al., 1999; Weiss & Prifitera, 1995).

The prediction of academic achievement remains a common practice in education as a means for guiding decisions related to student selection, diagnosis, and placement. At the same time, considerable debate exists regarding the accuracy of intelligence tests with members of different groups. The controversy reached its apex in the late 1970s and early 1980s when the concept of test bias came under critical scrutiny in terms of definitions, objective criteria, and empirical analyses (Jensen,

1980; Oakland, 1977; Oakland & Feigenbaum, 1979; Reynolds & Gutkin, 1980). This controversy remains today (Reynolds, 2000) despite prevailing evidence over the past 30 years demonstrating the cross-racial continuity of ability constructs and a general failure to uncover systematic differences in prediction between majority and minority groups (for a review, see Reynolds, Lowe, & Saenz, 1999).

However, in the atypical instances where differential prediction was detected, it most often operates against the majority group (i.e., criterion performance was generally overpredicted for minority groups). Findings of equitable validity hold for a variety of popular ability test scores, including Wechsler's various scales (Glutting, Oh, Ward, & Ward, 2000; Weiss & Prifitera, 1995; Weiss, Prifitera, & Roid, 1993), the Kaufman Assessment Battery for Children (Glutting, 1986), and the Developing Cognitive Abilities Test (Beggs & Mouw, 1980; Canivez, 1997, 1998; Canivez & Konold, 2001). For example, Weiss and Prifitera (1995) examined differential prediction of the Wechsler Individual Achievement Test (WIAT; Wechsler, 1992) scores with the WISC-III across race/ethnicity and gender. Results indicated that only 4 of the 12 simultaneous intercept and slope comparisons were statistically significant. No differences were observed between African American and White groups. One comparison (WIAT Reading) between Hispanic and White groups was statistically significant and the remaining three statistically significant simultaneous comparisons occurred between the female and male groups on WIAT Reading, Mathematics, and Writing. Follow-up analyses revealed that differences were limited to *y*-intercepts (and not slope) and that in all instances the effect sizes were small. Accordingly, no clinically meaningful or practical differences were indicated. (See Brown et al., 1999, for additional examples.)

Investigations of test score bias are codified in federal mandates outlined in the Individuals with Disabilities Education Improvement Act of 2004 (2004; Public Law [P.L.] 108-446), that continues "the longstanding requirement that procedures used for the evaluation and placement of children with disabilities not be discriminatory on racial or cultural basis" (p. 32). Moreover, examinations of potential test bias are endorsed by major professional organizations as reflected in the joint publication manual of *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). However, the rate at which new bias studies are appearing in the literature does not appear to match the rate at which new and revised instruments are being added to the field (Suzuki & Valencia, 1997). The decline is unfortunate when, in fact, more investigations are needed to address test bias with new assessments.

Given the common application of intelligence tests in forecasting achievement, criterion-related validity may be the most crucial form of validity evidence in relation to test bias (Reynolds et al., 1999). Criterion-related bias is indicated whenever errors in prediction vary as a function of group membership. Two types of errors are most likely. First, criterion-related bias is present when errors in prediction are constant across groups, that is, when *y*-intercept differences are present. Intercept differences indicate that one group is systematically overpredicted (or underpredicted) relative to

a common combined group regression line. Second, predictive bias is indicated when slope differences are obtained and errors in prediction vary across groups. In this second scenario, regression lines between the majority and minority groups are not parallel. Regardless of whether *y*-intercept or slope differences are present, the group with the higher criterion (e.g., achievement) score is underpredicted. Test scores can also show both *y*-intercept and slope bias in the form of either ordinal or disordinal group interactions, where the magnitude and direction of bias changes across points of the predictor scale. For instance, the group with the lower mean criterion score may be underpredicted when members obtain low scores on the predictor, but criterion performance may be overpredicted when its members obtain high scores on the predictor.

Accordingly, a thorough investigation of differential criterion-related validity involves joint consideration of both *y*-intercept and slope differences among groups. Potthoff's (1966) procedure is frequently used for addressing issues of differential prediction (Bossard, Reynolds, & Gutkin, 1980; Canivez & Konold, 2001; Glutting, 1986; Glutting, Oakland, & Konold, 1994; Shields, Konold, & Glutting, 2004; Weiss & Prifitera, 1995) because it provides a simultaneous *F* test for both *y*-intercept and slope differences, thereby, controlling Type I error rates. Following the identification of a statistically significant omnibus *F*, the procedure provides follow-up comparisons to detect whether differences exist between *y*-intercepts, slopes, or both. When either *y*-intercept or slope differences are found, the use of a common regression equation generally results in underpredicting criterion performance for the group with the higher mean criterion score. In the former case, errors of prediction are constant across all points of the predictor. However, slope differences suggest nonparallel regression lines and nonconstant errors of prediction; wherein, the size of the errors vary across different points of the predictor scale. Nonconstant errors of prediction are also problematic when both *y*-intercept and slope differences are observed. Here, however, interpretations become more challenging because the direction of bias may change across points of the predictor scale (see Reynolds & Kaiser, 1990).

To date, no investigation has examined differential criterion-related validity of scores from the WISC-IV and Wechsler Individual Achievement Test–Second Edition (WIAT-II; Wechsler, 2001). It is within this context that the current study examined whether the WISC-IV's Full Scale IQ (FSIQ) and General Ability Index (GAI) equitably predict concurrent WIAT-II achievement levels in Mathematics, Oral Language, Reading, Written Language, and total scores for groups differing by race/ethnicity, gender, and parent education level.

Method

Participants

Analyses were conducted on data obtained from the standardization linking sample of the WISC-IV (Wechsler, 2003) and the WIAT-II (Wechsler, 2001). Participants in the linking sample were selected to be nationally representative in accordance with

the 2000 U.S. Census variables of age, gender, race, geographic region, and parental education levels (Wechsler, 2003).

Participants ($N = 550$) ranged in age from 6 years, 0 months through 16 years, 11 months ($M_{age} = 11.58$, $SD_{age} = 3.22$) and included approximately equal numbers of males ($N = 282$) and females ($N = 268$). The three racial categories of Caucasian ($N = 334$), African American ($N = 86$), and Hispanic ($N = 101$) were sufficiently represented in the sample to serve as contrasting groups in the investigation of predictive bias. The remaining racial group representations of Asian Americans, Native Americans, and Others collectively comprised only 29 participants; too few to include as contrasting groups. Parent education levels included those with 8 to 11 years ($N = 102$), 12 years ($N = 145$), 13 to 15 years ($N = 172$), and 16 years ($N = 131$). Parent education level is a frequently used proxy for socioeconomic status in the development of standardized tests.

Instruments

The WISC-IV is a test of general intelligence and consists of 16 subtests (Ms = 10, SDs = 3), 10 of which are mandatory and contribute to measurement of four factor-based index scores: Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index. Each of the four indexes is expressed as a standard score (Ms = 100, SDs = 15). The FSIQ is composed of 10 subtests (3 verbal comprehension, 3 perceptual reasoning, 2 working memory, and 2 processing speed) whereas the GAI (Raiford, Rolfhus, Weiss, & Coalson, 2005) is composed only of the three verbal comprehension subtests and three perceptual reasoning subtests. The GAI is a global ability estimate not influenced by the lower *g* loading subtests that comprise the Working Memory Index and Processing Speed Index.

The WIAT-II is an individually administered clinical assessment of academic achievement and consists of nine subtests that can be combined to form four achievement composites: Mathematics, Oral Language, Reading, and Written Language. In addition, a total achievement score can be obtained. Each of the four composites and total score are expressed as a standard score (Ms = 100, SDs = 15).

Data Analysis

Potthoff's (1966) technique was employed to investigate criterion-related bias or differential prediction bias of WISC-IV scores across race, gender, and parent education level, by testing for constant errors in prediction (i.e., differences in regression slopes *and* *y*-intercepts). This procedure is superior to other methods of testing for slope and *y*-intercept differences because it reduces Type I errors through a single, simultaneous test of equivalence of both slopes *and* intercepts (Reynolds, 1982; Reynolds et al., 1999). As a result, it is viewed as the most efficient method for evaluating group differences in this context (Reynolds, 2000).

Standard scores from the WISC-IV (i.e., FSIQ and GAI) were used to predict WIAT-II achievement scores in Mathematics, Oral Language, Reading, Written Language,

and the WIAT-II total scale. Equality of slopes and *y*-intercepts were investigated across classifications of race (Caucasian vs. African American vs. Hispanic), gender (male vs. female), and parent education level (8-11 years vs. 12 years vs. 13-15 years vs. 16 years) as a proxy for socioeconomic status.

A statistically significant simultaneous *F* test shows if bias is present and indicates whether separate slope and intercept analyses are needed. Tests were simultaneous in terms of their inclusion of different sources of bias (i.e., slope and *y*-intercept) as well as being inclusive of the different groups within a given demographic (e.g., Caucasian vs. African American vs. Hispanic). Omnibus statistically significant simultaneous group comparisons were further examined to isolate the pairwise demographic groups that differed from one another (e.g., Caucasian vs. African American, Caucasian vs. Hispanic) and to determine whether the nature of the bias was related to slope and/or *y*-intercept differences. Given the relatively large number of statistical tests conducted with overlapping groups, Type I error rates were controlled through the use of a more conservative probability level ($\alpha = .01$) throughout. All Potthoff analyses were conducted with the MacPotthoff program (Watkins, 2005; Watkins & Hetrick, 1999).

Measures of effect size were calculated for statistically significant pairwise *y*-intercept and/or slope results by evaluating the predicted achievement score obtained from the minority groups' (i.e., African Americans, Hispanics, females, and lower parental education groups) regression line ($Y'g$) to the predicted achievement score that would be obtained from the use of a common regression line ($Y'c$) without regard to group membership. The difference between these two values ($Y'g - Y'c$) was divided by the criterion scale standard deviation (Sc) of the WIAT-II variable being predicted when bias was detected. This measure of effect size, $(Y'g - Y'c)/Sc$, is analogous to Cohen's *d* (1988). Positive values indicate that criterion scores for the minority group under consideration would be underpredicted if a common regression line were used. Likewise, negative values reflect instances in which the use of a common regression line would act to overpredict minority group criterion scores. Because this is the first examination of differential criterion-related validity of scores from the WISC-IV and WIAT, Cohen's (1988) benchmarks were used as proxies in the absence of more specific expectations regarding interpretation of the absolute values of the resulting coefficients, where .20 = small, .50 = medium, and .80 = large effect sizes. When follow-up tests revealed the presence of slope differences, the previously described measures of effect size were calculated at five points along the predictor scale (i.e., 70, 85, 100, 115, and 130) corresponding to WISC-IV's standard deviation intervals.

Results

Table 1 provides concurrent validity coefficients separately for the total sample and by demographic groups. Coefficients were large and statistically significant (ps $< .001$) between the investigated measures of the WISC-IV and WIAT-II. These trends were evident across all race, gender, and parental education level groups.

Table 1. Pearson Product–Moment Correlation Coefficients Between WISC-IV Predictors and WIAT-II Criteria for the Total Sample and Demographic Subgroups

	WIAT-II				
	Mathematics	Oral Language	Reading	Written Language	Total
WISC-IV FSIQ					
Total sample	.77	.75	.78	.77	.87
Race/ethnicity					
Caucasian	.70	.68	.70	.68	.80
African American	.80	.79	.77	.82	.89
Hispanic	.79	.78	.84	.84	.92
Gender					
Male	.78	.76	.79	.75	.87
Female	.76	.75	.75	.79	.87
Parent education (years)					
8-11	.82	.72	.76	.84	.89
12	.74	.70	.73	.79	.83
13-15	.66	.69	.70	.66	.80
16	.65	.67	.71	.63	.81
WISC-IV GAI					
Total sample	.74	.76	.75	.71	.84
Race/ethnicity					
Caucasian	.66	.68	.66	.59	.76
African American	.80	.79	.76	.82	.89
Hispanic	.74	.78	.81	.81	.88
Gender					
Male	.76	.77	.77	.72	.85
Female	.71	.76	.72	.72	.82
Parent education (years)					
8-11	.70	.73	.73	.80	.85
12	.68	.75	.72	.69	.79
13-15	.69	.69	.64	.60	.78
16	.60	.67	.68	.54	.75

Note: WISC-IV = Wechsler Intelligence Scale for Children–Fourth Edition; FSIQ = Full Scale IQ; GAI = General Ability Index; WIAT-II = Wechsler Individual Achievement Test–Second Edition. All correlations are statistically significant at $p < .0001$.

A total of 30 simultaneous demographic group comparisons were conducted between the WISC-IV's FSIQ and GAI, and the five WIAT-II achievement criteria by means of Potthoff's procedure. Table 2 presents *F* values, degrees of freedom, and corresponding *p* values for all simultaneous contrasts of slope and intercept differences between demographic groups. Five of the 30 simultaneous contrasts were statistically significant ($p < .01$). Four of the five statistically significant contrasts

Table 2. *F*, *df*, and *p* values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting WIAT-II Achievement Scores Using WISC-IV FSIQ and GAI Scores

	WIAT-II criteria				
Demographic comparisons	Mathematics	Oral Language	Reading	Written Language	total
WISC-IV FSIQ					
Caucasian vs. African American vs. Hispanic					
F	1.84	2.01	1.26	2.70	2.78
df	(4, 482)	(4, 479)	(4, 481)	(4, 475)	(4, 469)
p	.121	.092	.283	.030	.025
Males vs. females					
F	2.96	0.50	1.97	4.05	0.56
df	(2, 510)	(2, 506)	(2, 509)	(2, 503)	(2, 496)
p	.053	.609	.140	.018	.573
Parent education in years (8-11 vs. 12 vs. 13-15 vs. 16)					
F	3.05*	0.56	1.62	2.97*	2.59
df	(6, 506)	(6, 502)	(6, 505)	(6, 499)	(6, 492)
p	.006	.766	.140	.008	.018
WISC-IV GAI					
Caucasian vs. African American vs. Hispanic					
F	1.84	2.24	1.58	3.83*	2.79
df	(4, 495)	(4, 492)	(4, 494)	(4, 485)	(4, 479)
p	.120	.064	.179	.005	.026
Males vs. females					
F	0.56	3.00	1.02	8.48*	2.48
df	(2, 524)	(2, 520)	(2, 523)	(2, 514)	(2, 507)
p	.572	.051	.352	.0001	.085
Parent education in years (8-11 vs. 12 vs. 13-15 vs. 16)					
F	2.41	2.19	1.38	3.20*	2.33
df	(6, 520)	(6, 516)	(6, 519)	(6, 510)	(6, 503)
p	.027	.043	.221	.004	.032

Note: WISC-IV = Wechsler Intelligence Scale for Children–Fourth Edition; FSIQ = Full Scale IQ; GAI = General Ability Index; WIAT-II = Wechsler Individual Achievement Test–Second Edition. Data in parentheses are degrees of freedom.
*$p < .01$.

were related to the prediction of Written Language with the FSIQ (parent education, $p = .008$), and GAI (ethnicity, $p = .005$; gender, $p = .0001$; and parent education, $p = .004$). The fifth omnibus statistically significant contrast occurred for the prediction

of Mathematics with the FSIQ (parent education, $p = .006$). Follow-up evaluations of the statistically significant omnibus demographic group comparisons are described below.

Race

Simultaneous slope and *y*-intercept Potthoff comparisons between pairwise ethnicity groups revealed differential prediction of Written Language with the GAI for Caucasian versus Hispanic groups only, $F_{simultaneous}(2, 415) = 5.79$, $p = .003$. Wherein, differences were limited to slopes, $F_{slope}(1, 415) = 11.22$, $p = .0009$, and not *y*-intercepts, $F_{intercepts}(1, 416) = 0.35$, $p = .55$. Effect sizes for predicted Written Language scores between the Hispanic group equation and a common group equation were largely small across most points of the GAI scale (–.37, –.13, .11, .35, and .59). The exception occurred at the high end of the GAI scale, where the effect size was medium. Use of a common regression line resulted in a slight overprediction of Hispanic writing scores at the lower end (i.e., –2*SD* and –1*SD*) of the GAI scale and underprediction for higher GAI scores.

Gender

Subsequent evaluation of the statistically significant simultaneous Potthoff comparison between males and females for predicting Written Language from the GAI reported above revealed differences between *y*-intercepts, $F_{intercept}(1, 515) = 15.58$, $p = .0001$, but not slopes, $F_{slope}(1, 514) = 0.4$, $p = .528$. The standardized predicted difference based on females and a common regression equation, however, was notably small ($d = .13$).

Parent Education

Simultaneous slope and *y*-intercept Potthoff comparisons between pairwise parent education level groups revealed that the differential prediction of Written Language was limited to contrasts between groups with 8 to 11 years versus 16 years of parental education with both the FSIQ, $F_{simultaneous}(2, 209) = 4.73$, $p = .01$, and GAI, $F_{simultaneous}(2, 214) = 7.69$, $p = .0006$. In both instances, subsequent evaluation revealed that differences were limited to slope, FSIQ $F_{slope}(1, 209) = 8.54$, $p = .004$; GAI $F_{slope}(1, 214) = 13.39$, $p = .0003$, and not *y*-intercept, FSIQ $F_{intercept}(1, 210) = 0.89$, $p = .35$; GAI $F_{intercept}(1, 215) = 1.87$, $p = .17$, differences. Effect sizes for predicting Written Language with the FSIQ (–.20, –.06, .07, .21, .34) and GAI (–.28, –.09, .10, .30, .49) were generally small across most points of the predictor scale. In both instances, children of parents with lower education levels had overpredicted Written Language scores when they scored lower on the WISC-IV scales (–2SD and –1SD). In contrast, the use of a common regression line tended to underpredict children of parents with lower education levels when they scored higher on the WISC-IV scales.

Differential prediction of Mathematics from the FSIQ was also investigated through pairwise parent education level group contrasts. Results indicated that differences

were limited to contrasts between groups with 13 to 15 years versus 16 years of parental education, $F_{simultaneous}(2, 281) = 6.67, p = .002$. Subsequent evaluation of the simultaneous test revealed that differences were restricted to *y*-intercept, $F_{intercept}(1, 282) = 13.38, p = .0003$, and not slope, $F_{slope}(1, 281) = 0.00, p = .952$. The standardized predicted difference based on the 13- to 15-year parental education group and a common regression equation revealed slight favor for those with fewer years of parental education ($d = -.12$).

Discussion

Concurrent criterion-related correlations between the WISC-IV and WIAT-II were both large and statistically significant in the current study. When the total sample was considered, coefficients ranged from .75 to .87 for the FSIQ, and .71 to .84 for the GAI. Similarly, moderately large to large coefficients were also observed across subgroups defined by race/ethnicity, gender, and parent education.

Despite prevailing evidence over the past 30 years demonstrating the cross-racial continuity of ability constructs and a general failure to uncover systematic differences in prediction between majority and minority groups, considerable debate remains regarding the accuracy of intelligence tests with members of different groups (Reynolds, 2000). The primary purpose of the present study was to investigate differential errors in predication by the WISC-IV across groups differing by race/ethnicity, gender, and parent education. Whereas previous investigations with the earlier version of this instrument (i.e., WISC-III) examined differential criterion-related validity through consideration of race/ethnicity and gender (Weiss & Prifitera, 1995), the present study broadened the investigative framework to additionally include a proxy of socioeconomic status (i.e., parent education level). Potthoff's (1966) procedure was developed specifically for this purpose and was used to identify possible differential validity between groups.

In the aggregate, results indicated that differential relationships were observed in only 5 of the 30 omnibus comparisons. No differences were observed across race/ethnicity groups for WIAT-II and WISC-IV FSIQ scores, and only one statistically significant differential comparison was observed across race/ethnicity for *WISC-IV* GAI scores. The statistically significant difference was obtained between the Caucasian and Hispanic groups on WIAT-II Written Language. Here, however, only slope differences were observed in a slight overprediction of Hispanic Written Language scores at the lower end (i.e., –2*SD* and –1*SD*) of the GAI scale; and a slight underprediction for higher GAI scores. Effect sizes were generally small and not considered clinically meaningful until at the highest GAI level (+2*SD*) where the effect was somewhat larger.

Similarly, no differential prediction of WIAT-II scores were observed between gender groups for WISC-IV FSIQ scores. Only one statistically significant comparison was observed for the WISC-IV GAI. GAI prediction of WIAT-II Written Language scores demonstrated only *y*-intercept differences, and revealed only a small effect size. Clinically meaningful differences between boys and girls were not observed.

Differential prediction of WIAT-II Written Language scores across parent education was observed with the WISC-IV FSIQ and GAI between 8 to 11 years versus 16 years of parental education. Differences were limited to slopes and yielded small effect sizes. Wherein, children of parents with 8 to 11 years education had overpredicted Written Language scores when they scored lower on the WISC-IV FSIQ and GAI scales (–2*SD* and –1*SD*). In contrast, use of a common regression line tended to underpredict children of parents with 8- to 11-year education levels when they scored higher on the WISC-IV FSIQ and GAI scales. The final statistically significant differential WISC-IV prediction across parent education was obtained for the FSIQ predicting WIAT-II Mathematics scores between the 13 to 15 years versus 16 years of parental education groups. This difference was only in *y*-intercept and the difference again reflected a small effect size. As was the case with race/ethnicity and gender, the few parent education differences noted above yielded small effect sizes and were of limited clinical importance.

Results are consistent with previous research on the earlier version of the WISC (i.e., WISC-III) regarding race/ethnicity and differential prediction where differences were either not observed or when observed yielded small effect sizes (Weiss & Prifitera, 1995). Gender differences obtained in the present study were also similar to the Weiss and Prifitera (1995) results in which differences were noted only in *y*-intercept and demonstrated small effect sizes. Although no direct comparison for parent education can be made to the Weiss and Prifitera (1995) study given its lack of inclusion, the present results are consistent with findings on both the WISC-R (Hale, Raymond, & Gajar, 1982; Poteat, Wuensch, & Gregg, 1988) and other measures of children's intellectual abilities (Canivez & Konold, 2001).

There are, however, at least two factors that are worthy of mention as possible sources of influence on the results obtained in the current study. The first concerns the adequacy of the criterion used to evaluate predictive bias with the WISC-IV, and the second is related to the interpretation of results. The problem of identifying a suitable criterion variable that can be used to evaluate the validity and potential bias of a given predictor is well-known in the measurement literature (Crocker & Algina, 1986). In fact, some believe that the selection of a suitable criterion is often more difficult than selecting a good predictor (Nunnally & Bernstein, 1994). In the context of the current study, the WIAT-II was chosen as a viable criterion against which potential bias in the WISC-IV predictor was evaluated. From a practical standpoint, the decision to jointly investigate these instruments makes sense because educators and psychologists often may use both when evaluating children's educational proficiencies and disability. At the same time, it is possible that the WIAT-II operates differently for children from different groups and that the resulting scores do not have the same intended meaning in terms of the achievements the instrument was designed to measure. Thus, statistically significant bias (or nonbias) that was attributed to the WISC-IV, may in part be due to problems with the WIAT-II as a measure of student achievement to the extent that it is biased toward certain groups. Unfortunately, at the present time, the extent to which the psychometric properties of the WIAT-II differ for members of different groups is unknown.

The second factor that should be considered when evaluating the results of the current investigation relates to the notion of differential "prediction." Data for the present study were from the WISC-IV and WIAT-II standardization linking sample, and scores on the two measures were obtained at roughly the same point in time. Accordingly, the design is more "concurrent" than "predictive" in nature as there was no appreciable time delay between the administration of the WISC-IV and the WIAT-II such that WISC-IV scores were predicting future WIAT-II performance. Accordingly, results are best interpreted within the framework of a concurrent model but there remains a need for longitudinal investigation of differential "predictive" validity across these demographic variables.

Collectively, results from the current study support the use of the WISC-IV as a predictor of achievement. Little evidence suggested the presence of any meaningful bias. When findings of statistical and practical significance were considered jointly, results failed to indicate any meaningful group differences between (a) males versus females, (b) Caucasian versus African Americans, (c) Caucasian versus Hispanics, (d) African Americans versus Hispanics, (e) 8 to 11 years of parental education (PE) versus 12 years PE, (f) 8 to 11 years PE versus 13 to 15 years PE, (g) 8 to 11 years PE versus 16 years PE, (h) 12 years PE versus 13 to 15 years PE, (i) 12 years PE versus 16 years PE, and (j) 13 to 15 years PE versus 16 years PE. These outcomes are consistent with collective data obtained with intelligence tests across the past 30 years and provide more evidence for the equitable assessment of ability across groups differing by race/ethnicity, gender, and parental education level.

Authors' Note

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: Council on Faculty Research, Eastern Illinois University (2008 Summer Research Award to Gary L. Canivez).

References

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review*, *29*, 52-64.

Alfonso, V. C., & Pratt, S. I. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary*

intellectual assessment: Theories, tests, and issues (pp. 326-347). New York: Guilford Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beggs, D. L., & Mouw, J. T. (1980). *Developing Cognitive Abilities Test*. Carbondale: Southern Illinois University Press.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology*, *57*, 717-726.

Board of Scientific Affairs of the American Psychological Association. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101.

Bossard, M. D., Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the Stanford-Binet Intelligence Scale for Black and White children referred for psychological services. *Journal of Clinical Child Psychology*, *9*, 52-54.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484-502). New York: Guilford Press.

Brody, N. (2002). g and the one-many problem: Is one enough? In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *Novartis Foundation Symposium 233: The nature of intelligence* (pp. 122-135). New York: Wiley.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly*, *14*, 208-238.

Canivez, G. L. (1998, August). *Developing Cognitive Abilities Test (DCAT): Predictive Validity and Racial/Ethnic Bias*. Paper presented at the Focus on Science Session at the 1998 Annual Convention of the American Psychological Association, San Francisco, CA.

Canivez, G. L. (1997). *Developing Cognitive Ability Test (DCAT): Investigating psychometric characteristics and racial/ethnic bias*. Paper presented at the annual meeting of the American Psychological Association, Chicago.

Canivez, G. L., & Konold, T. R. (2001). Assessing differential prediction bias in the Developing Cognitive Abilities Test across gender, race/ethnicity, and socioeconomic groups. *Educational and Psychological Measurement*, *61*, 159-171.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace.

Flanagan, D. P., Andrews, T. J., & Genshaft, J. L. (1997). The functional utility of intelligence tests with special education populations. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 457-483). New York: Guilford Press.

Glutting, J. J. (1986). Potthoff bias analyses for K-ABC MPC and Nonverbal Scale IQs among Anglo, Black, and Puerto Rican kindergarten children. *Professional School Psychology*, *1*, 225-234.

Glutting, J. J., Oh, H. J., Ward, T., & Ward, S. (2000). Possible criterion-related bias of the WISC-III with a referral sample. *Journal of Psychoeducational Assessment*, *18*, 17-26.

Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to the Assessment of Test-Session Behavior for the WISC-III and WIAT: Possible race, gender, and SES effects. *Journal of School Psychology*, *32*, 355-369.

Hale, R. L., Raymond, M. R., & Gajar, A. H. (1982). Evaluating socioeconomic status bias in the WISC-R. *Journal of School Psychology*, *20*, 145-149.

Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review*, *21*, 271-284.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kaufman, A. S., & Lichtenberger, E. O. (2000). *Essentials of WISC–III and WPPSI–R assessment*. New York: Wiley.

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and achievement: Just how correlated are they? *Journal of Psychoeducational Assessment*, *21*, 244-260.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oakland, T. (Ed.). (1977). *Psychological and educational assessment of minority children*. New York: Brunner/Mazel.

Oakland, T., & Feigenbaum, D. (1979). Multiple sources of test bias on the WISC-R and Bender-Gestalt Test. *Journal of Consulting and Clinical Psychology*, *47*, 968-974.

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly*, *15*, 376-385.

Poteat, G. M., Wuensch, K. L., & Gregg, N. B. (1988). An investigation of differential prediction with the WISC-R. *Journal of School Psychology*, *26*, 59-68.

Potthoff, F. R. (1966). *Statistical aspects of the problem of bias in psychological tests* (Institute of Statistics Mimeo Series No. 479). Chapel Hill: University of North Carolina Department of Statistics.

Public Law (P.L.) 108-446. Individuals with Disabilities Education Improvement Act of 2004 (IDEIA). (20 U.S.C. 1400 et seq.). 34 CFR Parts 300 and 301. Assistance to States for the education of children with disabilities and preschool grants for children with disabilities; Final Rule. *Federal Register*, *71*(156), 46540-46845.

Raiford, S. E., Rolfhus, E., Weiss, L. G., & Coalson, D. (2005). General ability index (WISC-IV Tech. Rep. No. 4). San Antonio, TX: Psychological Corporation.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In F. Strickland & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249-285). New York: Plenum.

Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the WISC-R for Anglos and Chicanos referred to psychological services. *Journal of Abnormal Child Psychology*, *8*, 237-243.

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (pp. 847-525). New York: Wiley.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (3rd ed., pp. 549-595). New York: Wiley.

Shields, J., Konold, T. R., & Glutting, J. J. (2004). Validity of the Wide Range Intelligence Test: Differential effects across race/ethnicity, gender, and education level. *Journal of Psychoeducational Assessment*, *22*, 287-303.

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

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence. *American Psychologist*, *52*, 1103-1114.

Wasserman, J. D., & Tulsky, D. S. (2005). The origins of intellectual processing. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 3-38). New York: Guilford Press.

Watkins, M. W. (2005). MacPotthoff: Regression bias analysis [computer software]. University Park, PA: Pennsylvania State University/Author.

Watkins, M. W., & Hetrick, C. J. (1999). MacPotthoff: Automated calculation of the Potthoff regression bias procedure. *Behavior Research Methods, Instruments & Computers*, *31*, 710-711.

Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale*. New York: Psychological Corporation.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children*. New York: Psychological Corporation.

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

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

Wechsler, D. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2001). *Wechsler Individual Achievement Test–Second Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.

Weiss, L. G., & Prifitera, A. (1995). An evaluation of differential prediction of *WIAT* achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology*, *33*, 297-304.

Weiss, L. G., Prifitera, A., & Roid, G. (1993). The WISC-III and the fairness of predicting achievement across ethnic and gender groups. Advances in Psychoeducational Assessment; The Wechsler Intelligence Scale for Children: Third Edition. *Journal of Psychoeducational Assessment*, *Monograph Series*, 35-42.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, *25*, 9-23.