Journal of Psychoeducational Assessment
Volume 24 Number 4
December 2006 358-366
© 2006 Sage Publications
10.1177/0734282906289595
http://jpa.sagepub.com
hosted at
http://online.sagepub.com

Factorial Invariance of Woodcock-Johnson III Scores for African Americans and Caucasian Americans

Oliver W. Edwards
University of Central Florida

Thomas D. Oakland
University of Florida

Bias in testing has been of interest to psychologists and other test users since the origin of testing. New or revised tests often are subject to analyses that help examine the degree of bias in reference to group membership based on gender, language use, and race/ethnicity. The pervasive use of intelligence test data when making critical and, at times, life-changing decisions warrants the need by test developers and test users to examine possible test bias on new and recently revised intelligence tests. This study investigates factorial invariance and criterion-related validity of the Woodcock-Johnson III for African American and Caucasian American students. Data from this study suggest that although their mean scores differ, Woodcock-Johnson III scores have comparable meaning for both groups.

Keywords: *Woodcock Johnson; ethnic group differences; intelligence tests; achievement tests*

Despite changes in special education law (i.e., Public Law 108-446, Individuals With Disabilities Education Improvement Act [IDEIA]) that could limit the use of intelligence tests, data from these tests continue to be important in determining eligibility for special education placement (IDEIA, 2004) and in determining possible processing strengths and weaknesses. New or revised intelligence tests often are subject to analyses that help examine the degree of bias in reference to group membership based on gender, language use, and race/ethnicity (Scheuneman & Oakland, 1998). The findings from these studies suggest most frequently used intelligence tests are not statistically biased against ethnic groups (Brown, Reynolds, & Whitaker, 1999; Reynolds & Ramsey, 2003).

Bias in test use occurs "when deficiencies in the test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups" (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999, p. 74). Thus, bias occurs when the test's constructs or factors result in systematically different meanings across examinee subgroups.

Tests once were thought to be biased if scores from one group were higher or lower than scores from other groups, even if the factors were invariant across examinee subgroups. This standard for judging test bias is seriously flawed and has been replaced by an examination of forms of error, especially two possible contributors: systematic error (construct underrepresentation and construct-irrelevant components of test scores) and random error (AERA et al., 1999; see also Reynolds & Ramsey, 2003, for a comprehensive review of test bias).

Test developers attempt to eliminate or minimize all forms of error by taking steps to avoid construct underrepresentation and construct irrelevance. Moreover, test users generally are alert to the presence of random error and work to remove or reduce its presence (Frisby, 1999; Sattler, 2001). Despite test developers' best efforts to ensure their tests are psychometrically sound, further research of tests should be conducted by persons unaffiliated with the test developers.

Woodcock-Johnson III Tests of Cognitive Abilities (WJ III)

The WJ III (Woodcock, McGrew, & Mather, 2001) tests are used when making diagnostic and programming decisions (Bradley-Johnson, Morgan, & Nutkins, 2004). The WJ III Tests of Cognitive Abilities were designed to measure the intellectual abilities consistent with the Cattell-Horn-Carroll (CHC) theory of intelligence. CHC theory characterizes a hierarchical model of cognitive abilities through three strata. Stratum 1 abilities include the most specific or narrow abilities and may be measured by 40 or more tests. Stratum 2 abilities on the WJ III Standard Battery are measured by seven broad cognitive abilities: verbal comprehension (comprehension-knowledge, *Gc*), visual-auditory learning (long-term retrieval, *Glr*), spatial relations (visual-spatial thinking, *Gv*), sound blending (auditory processing, *Ga*), concept formation (fluid reasoning, *Gf*), visual matching (processing speed, *Gs*), and numbers reversed (short-term memory, *Gsm*). Stratum 3, the general factor, general intellectual ability (GIA or *g*), is derived by combining scores from Stratum 2 abilities (McGrew & Woodcock, 2001).

Data regarding the factor structure of the WJ III are reported for Caucasians and non-Caucasians but not specifically for African Americans (McGrew & Woodcock, 2001). The test authors report a root mean square error of approximation (RMSEA) fit statistic of .039 for Caucasian and non-Caucasians. According to the test authors, the aforementioned fit statistic suggests the WJ III measures the same constructs for Caucasians and non-Caucasians in the standardization sample. Additional research supports the invariance of WJ III subtests for ages across the life span (Taub & McGrew, 2004). However, neither the test manual nor additional research provides empirical evidence to indicate the subtests measure the same constructs for African Americans and Caucasian Americans explicitly. The reported analyses provide insufficient evidence of validity and thus require support from additional analyses that examine possible test bias for specific examinee subgroups (cf. Jensen, 1998).

Test Fairness

The WJ III were designed to minimize test bias associated with gender, race/ethnicity, or Hispanic origin (McGrew & Woodcock, 2001). Item development was conducted using recommended experts' views as to potential item bias and sensitivity. Items were modified or eliminated when statistical analyses supported an expert's assertion that an item was potentially unsuitable.

The CHC model is described as being uniform across ethnic groups (Carroll, 1993). Confirmatory factor analytic studies suggest the WJ III largely are invariant (i.e., very similar, although not identical) across examinees and reflect a fair formulation for examinee subgroups. Nonetheless, additional tests of invariance are needed to determine whether loadings for subtest scores and general intelligence differ specifically between African Americans and Caucasian Americans (McGrew & Woodcock, 2001). The test authors suggested additional research should be conducted to ascertain the degree of similarity among the WJ III's statistical properties for specific subgroups. Thus, additional analyses of data using the standardization sample were undertaken because the national database provides a large-scale representative sample of the U.S. population. In light of its large standardization sample and its reported oversampling of African Americans, the standardization sample provides the most comprehensive database to test the factorial invariance of the instrument.

The purpose of this study is to investigate the factorial invariance and criterion-related validity of the WJ III GIA factor for Caucasian Americans and African Americans in kindergarten through Grade 12. Examination of factor invariance is one of several different techniques to evaluate bias and determine whether tests measure the same constructs with similar fidelity across examinee subgroups (Watkins & Canivez, 2001). Although the WJ III use all three strata as part of its underlying framework, the GIA was emphasized because of its high reliability and validity for use when determining special education eligibility (Bradley-Johnson et al., 2004; McGrew & Woodcock, 2001). In addition, data from the seven cognitive abilities derived from the standard Cognitive Battery were analyzed primarily because these subtests are used most frequently by practitioners (cf. Bradley-Johnson et al., 2004; Sattler, 2001) and because of their direct contribution to GIA (McGrew & Woodcock, 2001).

Method

Participants

Participants include 1,978 Caucasian American and 401 African American students in kindergarten through Grade 12 who participated in norming the WJ III and for whom data were available. The sample was representative of the U.S. population with respect to gender, ethnicity, census region, and community size (McGrew & Woodcock, 2001). Detailed information regarding the demographic characters of the WJ III standardization sample and psychometric properties of the instrument are available in the technical manual (McGrew & Woodcock, 2001).

Statistical Procedures

Factor analyses were used to analyze the intercorrelations between the seven cognitive. Research using factor analysis suggests variables measured by cognitive abilities generally correlate highly (Jensen, 1998). The first unrotated factor is considered the general intelligence factor (*g*) and accounts for the largest proportion of the variance and highest *g* loading (Sattler, 2001). Principal components analysis and principal factor analysis may be used somewhat interchangeably to analyze the intercorrelations (Tabachnick & Fidell, 1996). Given the high reliability of the WJ III, principal factor analyses were used to identify the primary factor loadings (g loadings) for each of the seven cognitive abilities that comprise the standard battery. Principal factor analysis uses R^2 as the initial commonality estimate and provides a more accurate representation of the data (cf. Tabachnick & Fidell). The factor loadings for African Americans and Caucasian Americans were obtained separately.

The measurement invariance of data from the seven cognitive tests and *g* for the African American and Caucasian American samples was also examined using multiple group structural equation modeling (SEM). Multiple group SEM is used to establish measurement invariance because it controls for measurement errors and tests between-group invariance (cf. Jöreskog & Sörbom, 1999). LISREL (Jöreskog & Sörbom, 1999) was used to test the equivalence of the covariance structure and of the mean structure between the two groups.

The congruence coefficient, r_c, is used to measure agreement between factor structures across groups (Kamphaus, 2001). The congruence coefficient is an index of factor similarity and is interpreted similar to a Pearson correlation coefficient (Jensen, 1998).

> A value of r_c of +.90 is considered a high degree of factor similarity; a value greater than +.95 is generally interpreted as practical identity of the factors. The r_c is preferred to the Pearson *r* for comparing factors, because the r_c estimates the correlation between the factors themselves, whereas the Pearson *r* gives only the correlation between the two column vectors of factor loadings. (Jensen, 1998, p. 99)

The congruence coefficient has currency in measurement theory and is used to examine factor congruence or invariance (Reynolds & Ramsey, 2003). Thus, the congruence coefficient was used to measure agreement between the factor structures for African Americans and Caucasian Americans.

Criterion-related or convergent validity was also investigated. Pearson product moment correlations were used to determine relationships between general intelligence and 12 measures of achievement, including 9 academic achievement subtests and 3 broad achievement clusters from the WJ III: Tests of Achievement. The 9 achievement subtests form the three achievement clusters of broad reading, broad math, and broad written language. Correlation coefficients were examined for significance. The Fisher's *Z* transformation (cf. Cohen, 1988) was used to determine whether the correlation coefficients differed between African Americans and Caucasian Americans. This methodology analyzes the congruence in correlations for the two groups.

Taken together, the principal factor analysis, multiple-group SEM, r_c analysis, and convergent validity analysis provide important information about the dimensional and structural invariance of the WJ III across the two groups. These analyses are integral to testing the fidelity of the instrument.

Results

Factor Analysis

G loadings were obtained from principal factor analysis, using data from the seven cognitive ability tests (Table 1). The *g* loadings are sufficiently high for intelligence tests (Jensen, 1980) and, with one possible exception, are similar for both groups. The largest

Table 1
Factor Analysis of Standard Battery *g* Loadings for the Two Groups

WJ III Test	CHC Factor Model	CA[a]	AA[a]
Concept formation	Fluid reasoning - *Gf*	.78	.78
Verbal comprehension	Comprehension-knowledge - *Gc*	.76	.76
Visual-auditory learning	Long-term retrieval - *Glr*	.71	.67
Numbers reversed	Short-term memory - *Gsm*	.69	.59
Sound blending	Auditory processing - *Ga*	.59	.57
Visual matching	Processing speed - *Gs*	.55	.55
Spatial relations	Visual-spatial thinking - *Gv*	.47	.42

Note: CA = Caucasian Americans; AA = African Americans; WJ III = Woodcock-Johnson III: Tests of Cognitive Abilities; CHC = Cattell-Horn-Carroll.

a. First principal factor (maximum likelihood) *g* factor loadings calculated for the seven subtests, which are aligned with seven CHC broad factors.

Table 2
Mean IQs and Standard Deviations on the WJ III Standard Battery for Caucasian Americans and African Americans

WJ III Test	Mean IQ CA	Standard Deviation CA	Mean IQ AA	Standard Deviation AA
General intellectual ability	104.2	14.3	93.3	12.9
Sound blending	103.7	14.5	93.7	12.7
Verbal comprehension	103.3	14.1	90.9	12.7
Visual-auditory learning	103.2	14.0	98.5	13.2
Concept formation	103.2	14.2	94.2	14.2
Numbers reversed	101.8	15.1	97.1	17.3
Spatial relations	101.6	14.5	97.1	13.4
Visual matching	101.2	14.6	98.1	13.5

Note: CA = Caucasian Americans; AA = African Americans; WJ III = Woodcock-Johnson III: Tests of Cognitive Abilities.

between-group difference, .10, found on numbers reversed, suggests it is not as strong a measure of *g* for African Americans (.59) as for Caucasian Americans (.69). The factor loadings for the seven cognitive abilities are distributed across factors in such a manner that they closely approximate CHC theory. Tests with the highest *g* loadings generally assess reasoning, comprehension, deductive operations, and hypothesis-testing tasks (Carroll, 1993; Jensen, 1998; Valencia & Suzuki, 2001). The aforementioned holds true for these data; *Gf* and *Gc* exhibit the highest *g* loading, respectively.

Means and standard deviations for Caucasian Americans and African Americans are reported in Table 2. These data indicate mean scores for the two groups differ. Differences between the two groups are within the expected range given previous research findings of mean IQ differences between ethnic groups (cf. Sattler, 2001). The largest mean score difference is on *Gc*.

Because GIA may be more saturated with *Gc* than with other Stratum II abilities, it may be helpful to further investigate qualities measured by the verbal comprehension subtest. Overall, the data do not display restrictions in range or differences in the score distributions—qualities that could adversely influence subsequent analyses.

Multiple-group SEM was conducted to examine the equivalence of the covariance structure and of the mean structure. The comparative fit index (CFI), the normed fit index (NFI), and the non-normed fit index (NNFI) were used to test whether the model of covariance structural equality fits for the two groups (cf. Keith & Witta, 1997). These indices range in value from .00 to 1.00, and values > .95 suggest an outstanding fit, whereas values between .90 and .94 are considered an adequate fit (cf. Taub & McGrew, 2004). The RMSEA was also used as a measure of fit. RMSEA values range from .00 to 1.00, where 0 indicates a perfect fit. The results reveal a CFI of .99, a NFI of .98, a NNFI of .98, and a RMSEA of .059. The findings indicate that although chi-square was significant, $\chi^2 = 328.1$, $df = 8$, $p < .01$, all fit indices suggest an excellent fit for the spread of the data between the two groups.

When the means are constrained to equivalence ($\chi^2 = 500$, $df = 44$), the change in chi-square per degree of freedom change ($\Delta\chi^2 = 172$, $\Delta df = 36$) was statistically significant ($p < .01$), indicating the means were not equivalent. In this case, it was concluded that although the spread of the data provides an acceptable fit, the mean structure was not equal for the two groups.

Principal factor loadings for African Americans and Caucasian Americans are similar for each of the seven cognitive abilities. The congruence coefficient (r_c) analysis, conducted to determine whether *g* loadings are similar between African Americans and Caucasian Americans, is .99. Thus, the *g* factor structure is virtually identical for the two groups.

Correlations Between General Intelligence and Achievement

Correlation coefficients between GIA and three achievement clusters and nine achievement subtests are reported in Table 3. All correlations are significant ($p < .001$). The magnitude of correlations for the achievement clusters for reading, math, and written language averages .67 for both African Americans and Caucasian Americans. The magnitude of correlations for the nine achievement subtests averages .56 for African Americans and .54 for Caucasian Americans. Thus, correlations on the three achievement clusters and the nine achievement subtests are very similar for both groups. Fisher's *Z* transformations, used to compare correlations between GIA and the achievement clusters for reading, math, and written language as well as for each academic achievement test, indicate that correlations between general intelligence and the 12 academic achievement scores do not differ significantly by ethnicity (Table 4).

Discussion

Results from factor analysis, SEM, congruence coefficients, correlations coefficients, and Fisher's *Z* statistic are uniform in indicating the factor structure of the WJ III is consistent for African Americans and Caucasian Americans. The data show generally high

Table 3
Pearson Correlations Between General Intellectual Ability Standard Battery and Academic Achievement Scores for African Americans and Caucasian Americans

	Caucasian Americans			African Americans		
WJ III Test	*N*	GIA	*p*	*N*	GIA	*p*
Broad reading	1,851	.72	.001	365	.72	.001
Broad written language	1,788	.65	.001	357	.65	.001
Broad math	1,899	.64	.001	381	.64	.001
Letter-word identification	1,975	.62	.001	401	.65	.001
Reading fluency	1,851	.60	.001	365	.60	.001
Passage comprehension	1,975	.57	.001	400	.59	.001
Applied problems	1,974	.58	.001	399	.64	.001
Spelling	1,972	.55	.001	399	.54	.001
Writing fluency	1,789	.50	.001	357	.56	.001
Writing samples	1,953	.50	.001	396	.53	.001
Calculation	1,960	.49	.001	397	.48	.001
Math fluency	1,901	.46	.001	381	.49	.001

Note: GIA = general intellectual ability; WJ III = Woodcock-Johnson III: Tests of Cognitive Abilities.

Table 4
Fisher *Z* Transformation: *z* Test for Independent Correlations Between Caucasian Americans and African Americans for General Intelligence and Academic Achievement

	Caucasian Americans		African Americans		Both Groups
WJ III Test	Pearson's *r* General Intellectual Ability	Fisher *Z* Transformation	Pearson's *r* General Intellectual Ability	Fisher Z Transformation	*z* Scores
Reading: broad reading	.72	.91	.72	.91	.0000
Letter-word identification	.62	.72	.65	.77	–.0010
Reading fluency	.60	.70	.59	.69	.0003
Passage comprehension	.57	.66	.59	.68	–.0005
Math: broad math	.64	.76	.64	.76	.0000
Calculation	.49	.53	.48	.52	.0003
Math fluency	.46	.50	.49	.54	–.0008
Applied problems	.58	.66	.64	.77	–.0022
Written language: broad written language	.65	.77	.65	.78	.0004
Spelling	.55	.62	.54	.60	.0003
Writing fluency	.50	.55	.56	.63	–.0017
Writing samples	.51	.56	.53	.58	–.0004

Note: WJ III = Woodcock-Johnson III: Tests of Cognitive Abilities. All *z* scores are between $-1.96 < z_{\text{obs}} < 1.96$ and thus not significant.

and consistent *g* loading scores for African Americans and Caucasian Americans on the WJ III. The high congruence coefficient of .99 suggests the *g* factor structure is essentially identical for African Americans and Caucasian Americans. In addition, all fit indices are > .95, indicative of excellent fit and suggests covariant structural equivalence between the two groups. Although the mean IQs for the groups differ, the WJ III scores from the Cognitive Battery have comparable meaning for African American and Caucasian American students. Additionally, correlations between GIA and three achievement clusters and nine achievement subtests are similarly high and statistically significant for both groups.

These results are consistent with data from previous studies using confirmatory factor analysis and goodness-of-fit indices that revealed a comparable factor model, with the same factors and nearly identical directional patterns of factor loadings for most groups on the WJ III (McGrew & Woodcock, 2001). Furthermore, the collective findings from this and other studies using the WJ III provide some support for Carroll's (1993) assertion that CHC theory, one that forms the theoretical basis for the WJ III, is essentially invariant across racial/ethnic groups.

The largest between-group *g* loading difference, .10, was found on numbers reversed, a test of short-term memory. Among persons of similar IQs, African Americans generally obtain higher scores than Caucasian Americans on tests of short-term memory (Jensen, 1998). This may help account for the between-group difference found on numbers reversed.

The data from this study suggest that scores on WJ III have comparable meanings across two important examinee subgroups. As such, the instrument satisfies Standards 7.1 and 7.8 governing the fair use of tests with African American and Caucasian American students (AERA et al., 1999). According to these standards, test scores are considered to have comparable meanings across ethnic groups when scores generate similar inferences about examinees who are members of different ethnic groups. Given the current concerns about the use of standardized assessments and placement rates in special education for ethnic minorities, these findings provide some modest support for the integrity of the measure for use with the noted examinee subgroups. Thus, when using the WJ III Cognitive with African American and Caucasian American students, practitioners can be somewhat assured that possible score differences reflect differences in the underlying latent constructs rather than variations in the measurement operation itself (Watkins & Canivez, 2001).

Limitations

Two possible limitations moderate the implications of these findings. Data from each of the seven cognitive abilities were limited to one score. The employment of multiple scores may yield a different set of findings. This study chose to use one score from each of the seven cognitive abilities in light of the fact that many test users are limited by time and often decide to administer only the core or standard battery (cf. Sattler, 2001).

Correlations were analyzed to determine the probability that differences in the correlations observed between *g* and the achievement subtests were significant. However, analyses of correlations do not take into account possible differences in variable and factor variances. Despite factorial invariance, possible differences may affect test interpretation.

References

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

Bradley-Johnson, S., Morgan, S. K., & Nutkins, C. (2004). The Woodcock-Johnson Test of Achievement: Third edition. *Journal of Psychoeducational Assessment*, *22*, 261-274.

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

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

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

Frisby, C. L. (1999). Culture and test session behavior: Part I. *School Psychology Quarterly*, *14*, 263-280.

Individuals With Disabilities Education Act. (2004). Retrieved May 16, 2005, from http://www.ed.gov/offices/osers/idea/the_law.html

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

Jensen, A. R. (1998). *The* g *factor: the science of mental ability*. Westport, CT: Praeger.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence*. Needham Heights, MA: Allyn & Bacon.

Keith, T. A., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly*, *12*, 89-107.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock-Johnson III*. Itasca, IL: Riverside.

Reynolds, C. R., & Ramsey, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 67-93). New York: John Wiley.

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

Scheuneman, J. D., & Oakland, T. (1998). High-stakes testing in education. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 77-103). Washington, DC: American Psychological Association.

Tabachnick, B.G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly*, *19*, 72-87.

Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues*. Thousand Oaks, CA: Sage.

Watkins, M. W., & Canivez, G. L. (2001). Longitudinal factor structure of the WISC-III among students with disabilities. *Psychology in the School*, *38*, 291-298.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.