Psychology in the Schools, Vol. 38(4), 2001

LONGITUDINAL FACTOR STRUCTURE OF THE WISC-III AMONG STUDENTS WITH DISABILITIES

MARLEY W. WATKINS

The Pennsylvania State University

GARY L. CANIVEZ

Eastern Illinois University

If the factor structure of a test does not hold over time (i.e., is not invariant), then longitudinal comparisons of standing on the test are not meaningful. In the case of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), it is crucial that it exhibit longitudinal factorial invariance because it is widely used in high-stakes special education eligibility decisions. Accordingly, the present study analyzed the longitudinal factor structure of the WISC-III for both configural and metric invariance with a group of 177 students with disabilities tested, on average, 2.8 years apart. Equivalent factor loadings, factor variances, and factor covariances across the retest interval provided evidence of configural and metric invariance. It was concluded that the WISC-III was measuring the same constructs with equal fidelity across time which allows unequivocal interpretation of score differences as reflecting changes in underlying latent constructs rather than variations in the measurement operation itself. © 2001 John Wiley & Sons, Inc.

The Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) is one of the most widely used measures of intelligence (Stinnett, Havey, & Oehler-Stinnett, 1994). As such, it is an important component of the special education eligibility determination process for millions of students (Gresham & Witt, 1997). Given its popularity for such high-stakes decisions, it is critical that the WISC-III measure the same constructs across people and across time. Otherwise, WISC-III scores cannot be unambiguously interpreted (Hoyle, 2000).

It is generally assumed that the WISC-III measures the same construct(s) from person to person (Wechsler, 1991). To test this assumption, there have been many investigations of its factor structure among various groups of children. Initial analyses of the WISC-III standardization sample found that it was best represented by a four factor, first-order structure: (a) Verbal Comprehension (VC) composed of Information (IN), Similarities (SM), Vocabulary (VO), and Comprehension (CM) subtests; (b) Perceptual Organization (PO) composed of Picture Completion (PC), Picture Arrangement (PA), Block Design (BD), and Object Assembly (OA) subtests; (c) Freedom from Distractibility (FD) composed of Arithmetic (AR) and Digit Span (DS) subtests; and (d) Processing Speed (PS) composed of Coding (CD) and Symbol Search (SS) subtests (Wechsler, 1991). This four-factor solution was replicated in an independent nationally representative sample of 1,118 children (Roid, Prifitera, & Weiss, 1993) and among the Canadian normative sample (Roid & Worrall, 1997). Factor analytic results among special education populations have been somewhat inconsistent, but supportive of the major verbal and performance dimensions originally reported for the WISC-III normative sample (Konold, Kush, & Canivez, 1997; Kush, 1996; Poulson, 1995; Ravert & Watkins, 2000; Sullivan & Montoya, 1997). In general, the reported four factor structure of the WISC-III normative sample has been accepted with some disagreement surrounding the smaller FD and PS factors (Grice, Krohn, & Logerquist, 1999; Kush et al., in press).

Author Note: This research was supported, in part, by a Pennsylvania State University College of Education Alumni Society Faculty Research Initiation Grant and an Eastern Illinois University Faculty Development Grant.

The authors wish to express their gratitude to the school psychologists who generously responded to our request for WISC-III data. They also thank Tim Runge, Lisa Samuels, and Daniel Heupel for assistance in data entry.

Correspondence to: Marley W. Watkins, The Pennsylvania State University, Department of Educational and School Psychology and Special Education, 227 CEDAR Building, University Park, PA 16802. E-mail: mww10@psu.edu.

Intelligence is presumed to be an enduring trait; thus, factor analyses of tests measuring intelligence should also produce similar factor structures over time. Cross-sectional analyses of the WISC-III by age have been conducted to test this assumption. Sattler (1992) analyzed the WISC-III standardization sample across 11 separate age groups and reported that a three-factor (VC, PO, and PS) model best fit the normative data. In contrast, Keith and Witta's (1997) hierarchical confirmatory factor analysis of the WISC-III normative sample supported the primacy of a second-order *g* factor and four first-order factors (i.e., VC, PO, FD, and PS). Although there was agreement on three factors (i.e., VC, PO, and PS), there was disagreement on the samller FD factor and the utility of a higher-order structure.

There is, however, no empirical evidence regarding the stability of the factor structure of the WISC-III across time for the *same* individuals. This evidential lacuna is alarming because cross-sectional designs are inadequate tests of change over time (Willett, Singer, & Martin, 1998). Consequently, the present longitudinal study was conducted to investigate the temporal stability of the WISC-III factor structure among students with disabilities.

Method

Participants

Participants in the present study are a subset ($n = 177$ with data on 12 WISC-III subtests) of the total sample ($n = 667$) from a long-term WISC-III stability study by Canivez and Watkins (1998). Students included in the present study were independently classified with specific learning disability (SLD, $n = 115$), serious emotional disability (SED, $n = 9$), mental retardation (MR, $n = 17$), and other disabilities (i.e., speech, health, etc.; $n = 36$) by multidisciplinary evaluation teams according to state and federal guidelines governing special education classification.

Participants were students twice tested with the WISC-III. Of the 177 students who participated, 120 (67.8%) were male and 57 (32.2%) were female. Race/ethnicity included 146 (82.5%) Caucasian, 12 (6.8%) Hispanic/Latino, 16 (9.0%) Black/African American, 2 (1.1%) Native American/American Indian, and 1 (0.6%) Other/Missing. The mean age of students at first testing was 8.6 years ($SD = 1.76$) with a range from 6 to 13 years. The mean age of students at second testing was 11.4 ($SD = 1.84$) with a range from 7 to 16 years. The mean test-retest interval was 2.8 years ($SD = 0.54$) with a range of 1 to 4 years. Descriptive statistics for WISC-III subtest and composite scores across test and retest occasions are presented in Table 1.

Instrument

The WISC-III is an individually administered test of intelligence for children aged 6 years, 0 months through 16 years, 11 months. It contains 13 subtests, but only 10 are mandatory. The WISC-III was standardized on a nationally representative sample ($n = 2,200$) closely approximating the 1988 United States Census on gender, parent education (SES), race/ethnicity, and geographic region. Extensive evidence of reliability and validity is presented in the WISC-III Manual (Wechsler, 1991).

Procedure

Two thousand school psychologists were randomly selected from the National Association of School Psychologists membership list and invited to participate by providing test scores and demographic data for anonymous students who were administered the WISC-III during a special education triennial reevaluation. They were asked to report data if a student was administered the WISC-III during both the current reevaluation and an earlier evaluation, but there was no specification of how many cases to report nor were additional selection criteria (i.e., disability, gender,

Table 1
Means and Standard Deviations for WISC-III Subtest and Composite Scores Across Test and Retest Occasions (n = 177)

	Time 1		Time 2	
	M	*SD*	*M*	*SD*
Picture Completion	8.68	3.25	9.02	3.31
Information	7.99	3.13	7.97	3.18
Coding	8.34	3.31	7.39	2.85
Similarities	8.38	3.13	8.38	3.13
Picture Arrangement	8.59	3.42	8.69	3.69
Arithmetic	7.28	3.15	7.15	2.74
Block Design	8.40	3.41	8.23	3.68
Vocabulary	8.28	3.38	7.44	3.14
Object Assembly	8.61	3.24	8.59	3.57
Comprehension	8.82	3.67	8.54	3.61
Symbol Search	8.47	3.78	8.73	3.45
Digit Span	7.24	2.94	7.40	2.72
Verbal IQ	89.75	15.58	88.29	15.52
Performance IQ	91.10	15.56	90.40	16.31
Full Scale IQ	89.45	14.91	88.24	15.97

age, etc.) imposed. The 145 school psychologists from 33 states who responded provided an average of 4.6 cases each with a range of 1 to 25 cases. Additional details about this sample and procedure are provided in Canivez and Watkins (1998, 1999).

Cases were selected for inclusion in the present study if they contained complete data on all twelve WISC-III subtests that are associated with the four factors (Mazes was not included). Based on this criterion, 56 school psychologists from 26 states contributed WISC-III data on 177 evaluation-reevaluation cases. This reduction in usable cases is consistent with previous research which also found that practitioners do not routinely administer the Digit Span and Symbol Search supplemental subtests associated with the four factors of the WISC-III (Konold, Glutting, McDermott, Kush, & Watkins, 1999).

Analytic Procedures

Factorial invariance. There are several possible criteria to apply when testing whether the factor structure of the WISC-III is invariant over time. For example, the following tests, either alone or in combination, might be considered indicative of invariance: (a) equality of the covariance matrices, (b) equality of factor loadings, (c) equality of the number of common factors, (d) equality of factor variances, (e) equality of factor intercorrelations, or (f) equality of unique/error variances. Horn, McArdle, and Mason (1983) pointed out that these are *levels* of factorial invariance which may be successively more demanding.

In one sense, the most stringent test of factorial invariance is equivalence of covariance matrices. Horn and McArdle (1992) suggested that such precise equality is unlikely to hold in real-life situations and does not necessarily indicate that the same constructs are not being measured. Equivalence of factor and error variances are also very demanding tests of invariance and are generally not expected in applied research (Byrne, 1994; Keith et al., 1995; Long & Brekke, 1999).

However, the less stringent standard of factor loading equality, also called metric invariance, supports the assumption that the same constructs are being measured across time. This "provides support for a hypothesis of measurement invariance" and "is a reasonable ideal for research in the behavioral sciences" (Horn, 1991, p. 124). The least demanding test is of configural invariance. That is, equality of the number of salient factor loadings but not necessarily equivalence of the magnitude of those loadings. Configural invariance is the *minimum* condition required for factorial invariance (Schaie, Maitland, Willis, & Intrieri, 1998). Failure to achieve configural invariance suggests that changes have occurred in the factor structure and no interpretable comparisons of WISC-III scores could be made over time (Schaie et al., 1998). Although more stringent forms of invariance should be tested (Cunningham, 1991), only configural and metric invariance are necessary for unambiguous interpretation of the WISC-III.

Data analysis. Testing factorial invariance simultaneously across groups was first described by Joreskog (1971). This procedure entails a series of multiple-groups confirmatory factor analyses (CFA), beginning with one that restrains the covariance structure to be equal across groups. Failure to reject the hypothesis that the covariances are equal suggests that strong factorial invariance holds. However, rejection of this stringent model suggests that the groups are nonequivalent and successive CFA then test increasingly restrictive models to identify the source of noninvariance.

Unfortunately, there are problems with this classical approach to testing factorial invariance. As illustrated by Keith, Quirk, Schartzer, and Elliott (1999), this procedure requires several successive CFA. By conducting multiple tests, the overall Type I error rate is inflated (Bentler, 2000). Byrne (1994) also pointed out that this approach can generate contradictory results at successive levels.

These problems can be amelioriated by simultaneously testing the validity of equality constraints with Lagrange multiplier tests, which are asymptotically equivalent to chi-square difference tests (Anderson & Gerbing, 1988; Bentler, 1995). As implemented in the EQS program (Bentler & Wu, 1995), this multivariate strategy makes it unnecessary to compare a series of more restrictive models to determine factorial invariance because all equality constraints can be tested simultaneously in one CFA. Consequently, WISC-III factor invariance was analyzed via confirmatory factor analysis with EQS for the Macintosh version 5.6 (Bentler & Wu, 1995) according to the description provided by Byrne (1994).

Results

Determination of baseline models as a prerequisite for the testing of factorial invariance followed the CFA models presented in the WISC-III Manual (Wechsler, 1991) because the purpose of this study was to examine longitudinal factor invariance, not to establish the correct factor model for the WISC-III (Long & Brekke, 1999). Four alternative models were tested: Model One (One Factor) where all 12 subtests loaded on a general factor; Model Two (Two Factors) with 6 Verbal and 6 Performance subtests; Model Three (Three Factors) with 6 Verbal, 4 Performance, and 2 Processing Speed subtests; and Model Four (Four Factors) with 4 Verbal, 4 Performance, 2 Processing Speed, and 2 Freedeom from Distractability subtests.

Analysis of WISC-III subtests suggested that they followed a multivariate normal distribution (multivariate kurtosis = .44). Given multivariate normality and simple CFA models (Anderson & Gerbing, 1988; Bentler & Chou, 1987), maximum likelihood estimation was used (Byrne, 1994). WISC-III covariance matrices at Time 1 and Time 2 were analyzed separately to establish a unique baseline model for each (Byrne, 1994).

Statistical results indicated that Model Four was superior (see Table 2) at both times. The generalized likelihood chi-square statistic for Model Four was nonsignificant on both testing occa-

Table 2
Confirmatory Factor Analysis Fit Statistics for Disabled Sample on the WISC-II at Two Times

	χ^2			CFI[a]		RMSEA[b]	
Model	Time 1	Time 2	df	Time 1	Time 2	Time 1	Time 2
One	241.0**	218.1**	54	.815	.862	.140	.132
Two	100.8**	123.9**	53	.953	.940	.072	.087
Three	73.3*	71.0*	51	.978	.983	.050	.048
Four	62.2	63.4	48	.986	.987	.041	.043

[a]CFI = Comparative Fit Index, [b]RMSEA = Root Mean Squared Error of Approximation
*$p < .01$
**$p < .001$

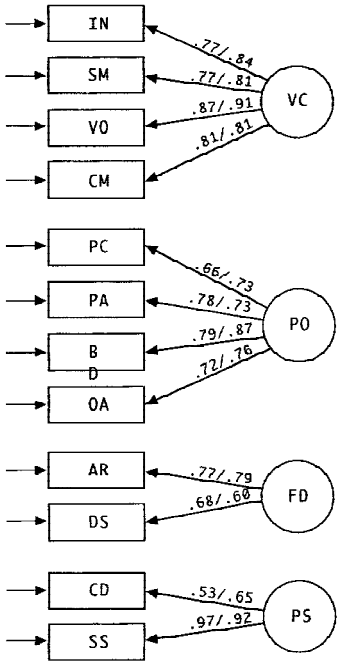
sions, indicating an acceptable fit to the sample data. Additionally, the chi-square difference test indicated statistically significant improvements in fit sequentially across models for both testing occasions (i.e., Model Three to Model Four difference chi-square for Time 1 $P = .011$; Time 2 $P = .055$). Third, the 90% confidence interval for the RMSEA statistic included zero only for Model Four, indicative of a very good fit (MacCallum, Browne, & Sugawara, 1996). Finally, Model Four met the combinational rule recommended by Hu and Bentler (1999) which requires both a CFI cutoff value close to .95 and an RMSEA value near .06 to minimize Type I and Type II error rates. These statistical results are also consistent with those reported for the WISC-III normative sample (Wechsler, 1991). Consequently, the first-order, four-factor model represented by Model Four was accepted as the baseline model for both test and retest occasions (see Figure 1 for this measurement model).

Although baseline models were equivalent, this does not guarantee equality across groups because estimation of baseline models involved no between-group constraints (Byrne, 1994). Consequently, test and retest data were next analyzed simultaneously with EQS (Bentler & Wu, 1995) to test for factorial invariance. Factor loadings, factor variances, factor covariances, and subtest error variances were constrained to be equal across time. Fit of this combined, restrained model was inferior to the baseline models with χ^2 (126) = 170, $P = .006$. More importantly, the multivariate Lagrange multiplier chi-square test was statistically significant, indicating that equality across the retest interval was not likely to be true. Three individual parameters significantly contributed to this multivariate effect: error variances for VO, CD, and AR.

When the model was respecified by releasing the cross-occasion constraint of equal error variances for VO, CD, and AR, model fit was significantly improved, χ^2 (123) = 148.5, $P = .058$. The resultant multivariate Lagrange multiplier chi-square test was not statistically significant, indicating that WISC-III factor loadings, factor variances, factor covariances, and subtest error variances (with the exception of VO, CD, and AR) were equivalent from Time 1 to Time 2 (or across test and retest). Thus, the WISC-III exhibited both configural and metric invariance across time.

Discussion

The longitudinal factor structure of the WISC-III was analyzed for invariance with a group of 177 students with disabilities tested, on average, 2.8 years apart. Equivalent factor loadings, factor variances, and factor covariances across the retest interval provided evidence of both configural



	PO	FD	PS
VC	.66/.75	.91/.90	.46/.52
PO		.69/.73	.66/.58
FD			.59/.59

FIGURE 1. Measurement model for WISC-III at Time 1/Time 2 for 177 students with disabilities tested an average of 2.8 years apart.

and metric invariance. Only three subtest error variances (AR, CD, and VO) were not equivalent across test and retest.

These longitudinal results are almost identical to the cross-sectional analyses reported by Keith and Witta (1997). However, the present study rejected the equality of error variances of three subtests whereas Keith and Witta accepted the equality of all subtest error variances. Nevertheless, error variances are generally not expected to be equal across groups or over time and their invariance does not invalidate equivalence of the factor structure (Byrne, 1994; Marsh, 1993). Therefore, these data suggest that the WISC-III measures the same constructs equally well across time and consequently allows unequivocal interpretation of score differences as reflecting changes in underlying latent constructs rather than variations in the measurement operation itself.

As with all research, these conclusions must be considered within the context of the limitations of this study. For example, non-random sample characteristics may make the results difficult to generalize. In this study school psychologists chose to report data from reevaluation cases that they personally selected and most did not administer all 12 WISC-III subtests. Additionally, the use of reevaluation cases meant that those students who were no longer enrolled in special education were not reevaluated and thus not included in the sample. Beyond sampling, there was no way to validate the accuracy of WISC-III test scores. Results could therefore have been influenced by administration, scoring, or reporting errors. Finally, the purpose of this study was to examine

longitudinal factor invariance, not to establish the correct factor model for the WISC-III. Consequently, the longitudinal invariance demonstrated in this study has not been strictly proven for alternative factor structures.

References

Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. Psychological Bulletin, 103, 411–423.

Bentler, P.M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P. M. (2000). Rites, wrongs, and gold in model testing. Structural Equation Modeling, 7, 82–91.

Bentler, P.M., & Chou, C. (1987). Practical issues in structural modeling. Sociological Methods and Research, 16, 78–117.

Bentler, P.M., & Wu, E.J.C. (1995). EQS for Macintosh user's guide. Encino, CA: Multivariate Software.

Byrne, B.M. (1994). Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming. Thousand Oaks, CA: Sage.

Canivez, G.L., & Watkins, M.W. (1998). Long term stability of the WISC-III. Psychological Assessment, 10, 285–291.

Canivez, G.L., & Watkins, M.W. (1999). Long term stability of the Wechsler Intelligence Scale for Children-Third Edition among demographic subgroups: Gender, race/ethnicity, and age. Journal of Psychoeducational Assessment, 17, 300–313.

Cunningham, W.R. (1991). Issues in factorial invariance. In L.M. Collins & J.L. Horn (Eds.), Best methods for the analysis of change (pp. 106–113). Washington, DC: American Psychological Association.

Gresham, F.M., & Witt, J.C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. School Psychology Quarterly, 12, 249–267.

Grice, J.W., Krohn, E.J., & Logerquist, S. (1999). Cross-validation of the WISC-III factor structure in two samples of children with learning disabilities. Journal of Psychoeducational Assessment, 17, 236–248.

Horn, J.L. (1991). Comments on "issues in factorial invariance". In L.M. Collins & J.L. Horn (Eds.), Best methods for the analysis of change (pp. 115–125). Washington, DC: American Psychological Association.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18, 117–144.

Horn, J.L., McArdle, J., & Mason, R. (1983). When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist, 1, 179–188.

Hoyle, R.H. (2000). Confirmatory factor analysis. In H.E.A. Tinsley & S.D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling (pp. 465–497). New York: Academic Press.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Joreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409–426.

Keith, T.Z., Fugate, M.H., DeGraff, M., Diamond, C.M., Shadrach, E.A., & Stevens, M.L. (1995). Using multi-sample confirmatory factor analysis to test for construct bias: An example using the K-ABC. Journal of Psychoeducational Assessment, 13, 347–364.

Keith, T.Z., Quirk, K.J., Schartzer, C., & Elliott, C.D. (1999). Construct bias in the Differential Ability Scales? Confirmatory and hierarchical factor structure across three ethnic groups. Journal of Psychoeducational Assessment, 17, 249–268.

Keith, T.Z., & 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.

Konold, T.R., Glutting, J.J., McDermott, P.A., Kush, J.C., & Watkins, M.W. (1999). Structure and diagnostic benefits of a normative subtest taxonomy developed from the WISC-III standardization sample. Journal of School Psychology, 37, 29–48.

Konold, T.R., Kush, J.C., & Canivez, G.L. (1997). Factor replication of the WISC-III in three independent samples of children receiving special education. Journal of Psychoeducational Assessment, 15, 123–137.

Kush, J.C. (1996). Factor structure of the WISC-III for students with learning disabilities. Journal of Psychoeducational Assessment, 14, 32–40.

Kush, J.C., Watkins, M.W., Ward, T.J., Ward, S.B., Canivez, G.L., & Worrall, F.C. (in press). Construct validity of the WISC-III for white and black students from the WISC-III standardization sample and for black students referred for psychological evaluation. School Psychology Review.

Long, J.D., & Brekke, J.S. (1999). Longitudinal factor structure of the Brief Psychiatric Rating Scale in Schizophrenia. Psychological Assessment, 11, 498–506.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1, 130–149.

Marsh, H.W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. American Educational Research Journal, 30, 841–860.

Poulson, K.M. (1995). Factor structure of the WISC-III for unclassified, learning-disabled, and high-IQ groups. Unpublished doctoral dissertation, Hofstra University.

Ravert, C.M., & Watkins, M.W. (2000, March). Meta-analysis of WISC-III factor analyses conducted with learning disabled students. Poster session presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA.

Roid, G.H., Prifitera, A., & Weiss, L.G. (1993). Replication of the WISC-III factor structure in an independent sample [Special issue, Monograph, WISC-III series]. Journal of Psychoeducational Assessment, 11, 6–21.

Roid, G.H., & Worrall, W. (1997). Replication of the Wechsler Intelligence Scale for Children-Third Edition four-factor model in the Canadian normative sample. Psychological Assessment, 9, 512–515.

Sattler, J.M. (1992). Assessment of children (3rd ed.). San Diego, CA: Jerome M. Sattler, Publisher.

Schaie, K.W., Maitland, S.B., Willis, S.L., & Intrieri, R.C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. Psychology and Aging, 13, 8–20.

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.

Sullivan, P.M., & Montoya, L.A. (1997). Factor analysis of the WISC-III with deaf and hard-of-hearing children. Psychological Assessment, 9, 317–321.

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

Willett, J.B., Singer, J.D., & Martin, N.C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. Development and Psychopathology, 10, 395–426.