

# Invariance of the Measurement Model Underlying the Wechsler Adult Intelligence Scale–III in the United States and Canada

Stephen C. Bowden

*University of Melbourne, Australia*

Rael T. Lange

*Riverview Hospital, Coquitlam, British Columbia, Canada*

Lawrence G. Weiss

*Harcourt Press Inc., San Antonio, Texas*

Donald Saklofske

*University of Calgary, Alberta, Canada*

A measurement model is invoked whenever a psychological interpretation is placed on test scores. When stated in detail, a measurement model provides a description of the numerical and theoretical relationship between observed scores and the corresponding latent variables or constructs. In this way, the hypothesis that similar meaning can be derived from a set of test scores can be tested by examination of a measurement model across groups. This study examines the invariance of a measurement model underlying Wechsler Adult Intelligence Scale–Third Edition scores in the U.S. and the Canadian standardization samples. The measurement model, involving four latent variables, satisfies the assumption of invariance across samples. Subtest scores also show similar reliability in both samples. However, slightly higher latent variable means are found in the Canadian normative sample.

**Keywords:** *intelligence; measurement invariance; Wechsler Adult Intelligence Scale–Third Edition; cross-cultural study*

Invariance in the measurement of psychological constructs across populations is a necessary condition to generalize results from validity studies to clinical practice. Therefore, establishing measurement invariance, or the equivalence of construct

---

**Authors' Note:** Special thanks to the Psychological Corporation, a Harcourt Assessment company, for permission to use the U.S. and Canadian WAIS-III standardization data. Wechsler Adult Intelligence Scale–Third Edition. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Data used with permission. All rights reserved. Please address correspondence to Stephen C. Bowden, Department of Psychology, University of Melbourne, Parkville, Victoria, Australia 3052; e-mail: [sbowden@unimelb.edu.au](mailto:sbowden@unimelb.edu.au).

measurement, has broad implications for the validity of clinical decisions, including meaningful interpretation of diagnostic inferences and classification decisions (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). In applied research, methods for examining comparability of construct measurement across groups are still little known (Keith, 1997; Vandenberg & Lance, 2000; Widaman & Reise, 1997).

Examining measurement invariance involves evaluation of the latent variable model underlying a set of test scores and testing for numerical equality across groups. Latent variables are the explicit definitions of psychological constructs (Byrne, Shavelson, & Muthén, 1989; Widaman & Reise, 1997). A full measurement model has been defined in terms of five matrices, derived from a mean-structure confirmatory factor analysis (CFA; see Byrne et al., 1989; Meredith, 1993; Reise, Widaman & Pugh, 1993; Vandenberg & Lance, 2000). The five matrices include (a) the matrix of factor loadings (the Lambda matrix:  $\Lambda$ ), (b) the vector of observed variable intercepts (tau:  $\tau$ ), (c) the matrix of residual variances (theta:  $\theta$ ), (d) the vectors of means of the latent variables in each group (alpha:  $\alpha$ ), and (e) the matrix of variances and covariances between factors or latent variables (psi:  $\psi$ ). The first three matrices are usually termed the measurement components of the model because these matrices define the measurement of the latent variables in terms of the observed variables in each group, and the reliability of the measurement through the observed scores (see Vandenberg & Lance, 2000; Widaman & Reise, 1997). The final two matrices, (d) and (e) above, usually termed the structural components, represent the values of and relationships between the latent variables.

The numerical values of the first three matrices in the measurement model, (a) to (c) above, have critical meaning when establishing the invariance of a measurement model across groups. Differences in the regression relationship relating the latent variables to the observed scores, provided by the numerical values of the factor loading and observed variable intercept matrices, may imply that the latent variables are measured on a different scale in different groups (Byrne et al., 1989; Horn & McArdle, 1992). Such a finding may seriously hamper interpretation of test scores and generalization of validity studies across groups (Taub, McGrew, & Witta, 2004). As noted by Widaman and Reise (1997), “for test scores to be comparable across ostensibly distinct examinee populations, the observed test items, or indicators, must have identical, or invariant, quantitative relationships with the latent variable for each population of interest” (p. 282).

Most interpretation of psychological tests in applied settings, other than the population in which the test was developed, involves the assumption of measurement invariance. Without this assumption the scientific basis of assessment is uncertain (Horn & McArdle, 1992). As noted above there is, to date, little research bearing on the measurement invariance of commonly used tests. The purpose of this study was to examine, in detail, the measurement invariance of the latent variable model derived from the U.S. standardization sample for the WAIS-III (The Psychological

Corporation, 2002), with the model derived from the Canadian WAIS-III standardization (The Psychological Corporation, 2001). The demonstration of measurement invariance is necessary for the inference that the test measures the same psychological constructs on the same scales in different populations. On the assumption of measurement invariance, the equivalence of latent variables in the U.S. versus Canadian normative samples can then be examined.

## Method

### Participants

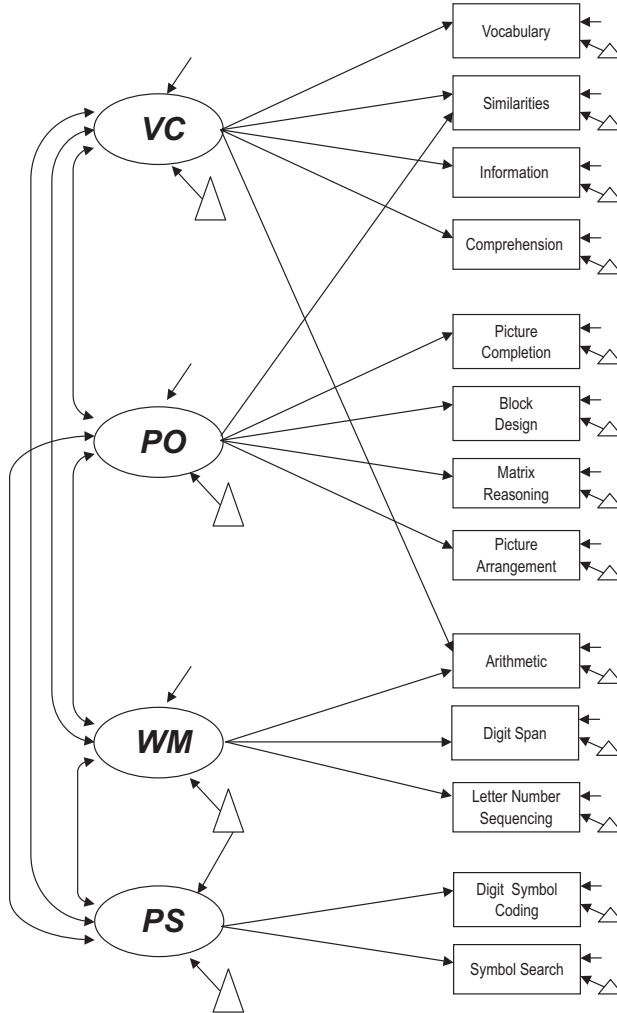
Participants were drawn from the U.S. WAIS-III standardization sample ( $n = 2,450$ ) and the Canadian WAIS-III standardization sample ( $n = 1,105$ ), obtained with permission from the Psychological Corporation, a Harcourt Assessment company.

The U.S. WAIS-III standardization sample was selected to match the demographic characteristics of the 1995 U.S. Census and is divided into 13 age groups each containing 200 individuals, with the exception of the two oldest age groups (i.e., 80-84 and 85-89), which had 150 and 100 individuals, respectively. The U.S. norms are generated using the customary method of forced normalization (i.e., the norms for each age group are based on a subset of the entire sample). Full details regarding the demographic characteristics and inclusion criteria of the sample can be found in the WAIS-III/WMS-III technical manual (The Psychological Corporation, 2002) and reviewed by Tulsky and Ledbetter (2000).

The Canadian WAIS-III standardization sample was selected to match the demographic characteristics of the 1991 Canadian Census and is divided into 13 age groups ranging from age 16 to 89 years. The Canadian norms were derived from people who were fluent in English, although 30% of the sample identified themselves as ethnically French or other European (see Table 2.3 of The Psychological Corporation, 2001). The Canadian norms were generated using a statistical technique known as continuous norming (i.e., the norms are based on the entire sample). Full details regarding the demographic characteristics and inclusion criteria of this sample can be found in the WAIS-III Canadian technical manual (The Psychological Corporation, 2001).

Below we analyze two samples. Initially, only participants in both samples who completed all 13 subtests of the test were included (for detail of subtests see Figure 1). We were interested to examine this subsample because the 13-subtest version of the WAIS-III, including Letter-Number Sequencing provides the strongest basis for the empirically derived factor scores (Tulsky & Price, 2003). Therefore, a subsample of 1,299 individuals from the U.S. standardization sample and a subsample of 886 individuals from the Canadian standardization sample were included in the first invariance analysis. Demographic information for these subsamples is presented in Table 1.

**Figure 1**  
**Four-Factor Model Identified as Best-Fitting**  
**in the U.S. and Canadian Normative Samples**



Note: This model was used as the basis for testing measurement equivalence across samples (see Table 2 and text for details). In line with convention, rectangles represent observed variables, ellipses represent latent variables, and triangles represent intercepts of observed variables or means of latent variables. Double-headed arrows represent covariances between latent variables, and single-headed arrows represent regressions of observed variables on latent variables, variances of latent variables, or residual variances of observed variables.

**Table 1**  
**Demographic Characteristics of the 13-Subtest Subsamples**  
**From the U.S. and Canadian WAIS-III Standardization Data**

	United States	Canada
Overall <i>ns</i>	1,299	866
No. of females / no. of males	696/603	506/380
% Female / % male	54%/46%	57%/43%
Mean age ( <i>SD</i> ), range	48.6 (23.8), 16-90	42.4 (21.7), 16-88
Education, <i>n</i> (%)		
8 years or less	149 (11.5%)	83 (9.4%)
9-11 years	164 (12.6%)	181 (20.4%)
Completed high school	457 (35.2%)	139 (15.7%)
College or vocational	284 (21.9%)	319 (36%)
University degree	245 (18.9%)	164 (18.5%)

Note: All data are derived from the U.S. and Canadian standardization samples of the *Wechsler Adult Intelligence Scale—Third Edition*. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Used with Permission. All rights reserved.

Subsequently, the 12-subtest version was examined across the entire, census-matched samples.

## Results

### Thirteen-Subtest Data Sets

*Data analysis and baseline model estimation.* Maximum likelihood CFA was conducted with Mplus 3.12 (Muthén & Muthén, 2004). The basis for analysis was the variance–covariance matrix of WAIS-III subtest total raw scores in both samples. A model of WAIS-III scores comprising four latent variables is well established (The Psychological Corporation, 2002; Tulsy & Price, 2003) and has been shown to be invariant across the entire age range of the U.S. standardization sample (Bowden, Weiss, Holdnack, & Lloyd, 2006). These latent variables are Verbal Comprehension (VC), Perceptual Organization (PO), Working Memory (WM), and Processing Speed (PS). Because the four-factor model is well established, relative model fit in the U.S. standardization sample will not be reexamined here. However, it should be noted that recent analyses indicate improved model fit if Arithmetic is allowed to load on both VC and WM, and Similarities loads on both VC and PO (see Figure 1 and Bowden, Cook, Bardenhagen, Shores, & Carstairs, 2004; Bowden et al., 2006).

Similarly, in the Canadian sample, the four-factor model with cross loadings for Arithmetic and Similarities (see Figure 1) also provided the best fit— $\chi^2(n = 886;$

$df=57$ ) = 215.67;  $p < .01$ —replicating the similar four-factor solution originally reported from the Canadian standardization sample (Saklofske, Hildebrand, & Gorsuch, 2000). Again, the relative fit of factor models has been examined extensively in the Canadian standardization and will not be reiterated here (The Psychological Corporation, 2001; Saklofske et al., 2000). However it is important to note that in the present analysis, the four-factor model with cross loadings for Arithmetic and Similarities provided a statistically significantly better fit than a simpler three-factor model comprising VC, PO, and WM, with this latter factor also incorporating Digit-Symbol-Coding and Symbol Search,  $\chi^2(n=886; df=62) = 752.00$ ;  $p < .01$ . Similarly, the four-factor model with cross loadings provided better fit than a five-factor model including a Quantitative Ability factor in addition to VC, PO, WM, and PS,  $\chi^2(n=886; df=56) = 327.48$ ;  $p < .01$ . The pattern of results for the Canadian sample is similar to that obtained for the U.S. standardization sample (Bowden et al., 2006; The Psychological Corporation, 2002) and to that obtained in analysis of the full Canadian standardization sample (The Psychological Corporation, 2001) although the four-factor model with cross loadings has not been examined previously.

To determine whether nonnormal data was producing misestimation of model fit, the above relative model fit comparisons were repeated using the mean-adjusted maximum likelihood (MLM) estimator, which is robust to departures from normality (Muthén & Muthén, 2004). A very similar pattern of model fit was obtained with the MLM estimator in both samples, suggesting that nonnormality did not lead to inappropriate conclusions regarding relative or absolute model fit.

*Measurement invariance.* Finding that the same baseline CFA model provides the best fit to the data in both samples has been termed *configural* invariance (Widaman & Reise, 1997). In other words, the same configuration of indicator variables or observed scores measures the same number of latent variables in each sample. Configural invariance provides the starting point for examination of numerical or metric invariance of the measurement model across groups.

Following the strategy outlined by Widaman and Reise (1997), goodness-of-fit statistics were examined against the baseline model when estimated simultaneously in both groups (see the baseline Invariance Model 1 in Table 2). As detailed below, stepwise invariance or equality constraints were then placed on the factor loading, observed score intercept, and residual matrices, in order, and the model reestimated at each step to evaluate the equality of these components across groups. This approach uses a particular type of CFA, known as a mean-structure analysis, that estimates all five of the matrices described above in both samples simultaneously, including the latent variable means, variances, and covariances in each group (Muthén & Muthén, 2004). The covariance matrix of raw scores in both samples was used as the basis of the analysis.

A model specification approach identical to that used in previous studies was followed (Bowden et al., 2004; Bowden et al., 2006; Widaman & Reise, 1997). Initially,

**Table 2**  
**Summary of Tests for Metric Invariance of Intelligence Measurement in 13 WAIS-III Subtest**  
**Raw Scores Across the U.S. ( $n = 1,299$ ) and Canadian ( $n = 886$ ) Standardization Samples**

Invariance Model	$\chi^2$	df	CFI	TLI	RMSEA		SRMR	ECVI		Gamma 1	
					Lower/Upper Bounds (95%)	Lower/Upper Bounds (95%)		Lower/Upper Bounds (95%)	Lower/Upper Bounds (95%)		
Model 1	522.209 *	114	.979	.972	.0573	.024	.3264	.2897	.3677	.9661	.9774
Baseline (configural) invariance					.0514					.9703	
Model 2	558.620 *	125	.978	.973	.0564	.031	.2949	.3329	.3755	.9643	.9758
Model 1 and all factor loadings invariant					.0507	.032	.3520			.9670	
Model 3	618.650 *	134	.975	.971	.0576	.035	.3119	.3709	.3968	.9606	.9728
Model 2 and all observed intercepts invariant					.0521					.9634	
Model 4	686.234 *	147	.973	.971	.0580	.032	.3285	.3285	.4179	.9568	.9695
Model 3 and all residual variances invariant					.0528						

Note:  $p < .01$  for the  $\chi^2$  test. CFI = comparative fit index; TLI = Tucker-Lewis index or nonnormed fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; ECVI = expected cross-validation index. All data are derived from the U.S. and Canadian standardization samples of the Wechsler Adult Intelligence Scale-Third Edition. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Used with permission. All rights reserved.

factor variances were fixed to unity in the factor variance–covariance matrix for the U.S. sample, and freely estimated in the Canadian sample. Values in the vector of latent factor means were fixed to zero in the U.S. sample and freely estimated in the Canadian sample. In addition, the first factor loading for each factor in the  $\Lambda$  matrix and the corresponding observed-variable intercept in the  $\tau$  matrix, respectively, were constrained to equality (but not unity) across groups to identify the model.

A variety of statistics and fit indices were used to evaluate relative model fit. Apart from  $\chi^2$ , other fit indices examined were the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the standardized root mean square residual (SRMR). We also examined the multiple group versions of the root mean square error of approximation (RMSEA), Gamma 1 (a multiple group version of the goodness-of-fit index), and the expected cross-validation index (ECVI). For these last three indices, it is now possible to calculate confidence intervals, particularly useful for comparison of nested models (Dudgeon, 2004). Details of these fit indices can be found in a variety of sources (e.g., Byrne, 1998; Dudgeon, 2004; Muthén & Muthén, 2004).

Small sample statistics such as  $\chi^2$  may be overly sensitive to changes in goodness of fit when applied to large samples (see Cheung & Rensvold, 2002). Therefore, to evaluate change in fit with stepwise invariance restrictions, we placed most emphasis on the overall pattern of fit (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000) and those indices for which confidence intervals can be estimated, namely the RMSEA, ECVI, and Gamma 1. In their review, Vandenberg and Lance (2000) suggested that when comparing invariance models across groups, absolute RMSEA values below .06 and SRMR values below .08 reflect excellent fit, and stepwise changes in the CFI of more than  $-.02$  represent definite loss of fit. All of the fit statistics were estimated using Mplus 3.12 (Muthén & Muthén, 2004), except for the multigroup RMSEA, ECVI, and Gamma 1 that were calculated using a program provided by Dudgeon (2004).

As noted, the results for the baseline model when estimated simultaneously in both groups in the mean-structure CFA are shown in Table 2 as Invariance Model 1. The  $\chi^2$  value and degrees of freedom for this model are simply the sum of the values from each model estimated in the separate samples. All of the other fit indices for Invariance Model 1 in Table 2 indicate that this baseline model provides an excellent fit to the model of subtest raw scores.

In the next step, we tested the hypothesis that the numerical values in the  $\Lambda$  or factor loading matrices were equal across groups. The results are shown as Invariance Model 2 in Table 2. Although a statistically significant increment in  $\chi^2$  suggests statistically significant loss of fit compared to the baseline model (invariance Model 1 in Table 2), other fit statistics suggest that invariance Model 2 shows no appreciable loss of fit and in fact continues to provide an excellent fit to the data in both samples. In absolute terms the CFI, TLI, and SRMR all suggest excellent fit and no obvious change from the values obtained for Invariance Model 1 in terms of the criteria described by Vandenberg and Lance (2000). In addition, the values of

the RMSEA, ECVI, and Gamma 1 were all within the 95% confidence intervals for the same indices estimated for Invariance Model 1. Therefore, the assumption of invariant factor loadings across groups was retained.

To evaluate the invariance of the numerical values of the observed variable intercepts in the  $\tau$  vector, these values were then held constant across groups in addition to the factor loadings. The results are shown as Invariance Model 3 in Table 2. The results again show a statistically significant increment in  $\chi^2$  but the other statistics show little change, and the values of the RMSEA, ECVI, and Gamma 1 indices were again within the confidence intervals for the respective indices estimated for the preceding step. Overall, the fit statistics indicate excellent fit for Invariance Model 3, so the assumption of invariance of observed variable intercepts across groups was retained. This level of measurement invariance is known as scalar or strong metric invariance (Widaman & Reise, 1997).

Finally, invariance of the residual variances in the  $\theta$  matrix was examined by placing additional equality constraints on these elements. Results can be seen as Invariance Model 4 in Table 2 and again, apart from a statistically significant change in  $\chi^2$ , there was little absolute change in the other fit indices, and the RMSEA, ECVI, and Gamma 1 fell within the confidence intervals for the preceding step. These results prompt retention of the assumption of equality of the residual variances or reliability of observed scores across the U.S. and Canadian normative samples.

## **Twelve-Subtest Data Sets**

*Measurement invariance.* To examine measurement invariance in representative samples, the above analyses were repeated in the full U.S. ( $n = 2,450$ ) and Canadian ( $n = 1,105$ ) standardization samples, comprising data for 12 subtests, excluding Letter-Number Sequencing. Thirteen cases in the Canadian sample were excluded because of missing data on Digit Symbol Coding, resulting in a final sample of 1,092. Although results are not shown, relative baseline model fit was examined in the Canadian sample as reported above for the smaller 13-subtest sample. This process led to the same conclusion of preference for the modified four-factor model (as shown in Figure 1 minus the Letter-Number Sequencing subtest). Reexamination of relative model fit using the MLM estimator again suggested that departures from normality were not influencing absolute model fit, or choice of model.

The invariance analysis was then repeated, and fit statistics and indices for the stepwise restrictions on invariance testing are shown in Table 3. The method of analysis and sequence of steps is identical to that reported above for the 13-subtest samples. Briefly, inspection of the fit statistics and indices at each step in Table 3 led to the same conclusions regarding retention of the null hypothesis of invariance as was the case for the 13-subtest samples reported above (see Invariance Models 1 to 4 in Table 3).

**Table 3**  
**Summary of Tests for Metric Invariance of Intelligence Measurement in the Representative 12-Subtest**  
**Version of WAIS-III Raw Scores Across the U.S. ( $n = 2,450$ ) and Canadian ( $n = 1,092$ ) Standardization Samples**

Invariance Model	$\chi^2$	df	CFI	TLI	RMSEA		SRMR	ECVI		Gamma I	
					Lower/Upper Bounds (95%)	Lower/Upper Bounds (95%)		Lower/Upper Bounds (95%)	Lower/Upper Bounds (95%)		
Model 1	626.957*	92	.983	.975	.0573	.0573	.019	.2272	.2272	.9754	.9754
Baseline (configural) invariance					.0523	.0624		.2018	.2553	.9710	.9795
Model 2	694.284*	102	.981	.975	.0573	.0573	.030	.2405	.2405	.9729	.9729
Model 1 and all factor loadings invariant					.0525	.0621		.2137	.2701	.9682	.9771
Model 3	774.171*	110	.978	.974	.0584	.0584	.032	.2585	.2585	.9697	.9697
Model 2 and all observed intercepts invariant					.0538	.0631		.2301	.2897	.9648	.9741
Model 4	843.355*	122	.977	.975	.0578	.0578	.034	.2712	.2712	.9672	.9672
Model 3 and all residual variances invariant					.0534	.0622		.2415	.3037	.9621	.9718
Model 5	904.004*	132	.975	.975	.0575	.0575	.053	.2827	.2827	.9649	.9649
Model 4 and all latent variable variances and covariances invariant.					.0533	.0617		.2519	.3162	.9598	.9697
Model 6	1071.490*	126	.969	.968	.0651	.0651	.090	.3334	.3334	.9574	.9574
Model 4 and all latent variable means invariant					.0609	.0694		.2995	.3701	.9518	.9626

Note:  $p < .01$  for  $\chi^2$  test. For abbreviations, see Table 2. All data are derived from the U.S. and Canadian standardization samples of the Wechsler Adult Intelligence Scale-Third Edition. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Used with permission. All rights reserved.

*Latent variances and covariances.* In the context of an invariant measurement model across groups, the values of the latent variables were examined. The fully invariant measurement model (Invariance Model 4 in Table 3) was used as the basis for estimation of latent variable variances and covariances with equality restraints across groups. Invariance Model 5 in Table 3 reports the results of equality constraints imposed on the latent variable variances and covariances. Although there is a statistically significant increment in  $\chi^2$  compared to Invariance Model 4 in Table 3, none of the fit indices suggest appreciable loss of fit.

Parameter estimates from the model with no equality constraints on latent variable variances or covariances are shown in Table 4. In the upper part of Table 4 (shown on the diagonal in bold) are the latent variable variances, fixed to 1 for model identification in the U.S. sample and freely estimated in the Canadian sample. The variances for VC and WM in the Canadian sample were both less than 1 standard error different from unity. The variances for PO and PS were both between 2 and 3 standard errors less than the unit value in the U.S. sample.

Below the diagonal in the upper part of the Table 4 are shown the covariances between latent variables freely estimated in both samples. Some of the covariances between latent variables differed between samples by more than 2 standard errors. Inspection of modification indices for Invariance Model 5 in Table 3 showed that if the equality constraints for the covariances for VC with PO, VC with PS, and PO with WM, respectively, were relaxed then there was no longer a statistically significant increment in  $\chi^2$  between Invariance Models 4 and 5 (Table 3).

*Latent means.* In the lower part of Table 4 are the latent variable means, fixed to 0 in the U.S. sample and freely estimated in the Canadian sample. Testing equality of the latent means across groups also led to rejection of the assumption of invariance (Invariance Model 6 in Table 3). Not only was there a statistically significant increment in  $\chi^2$  compared to Invariance Model 4, but all of the fit indices for which confidence intervals could be estimated also led to rejection of the assumption of no loss of fit between Invariance Models 4 and 6 (Table 3). Inspection of the modification indices for Invariance Model 6 showed that it was necessary to relax equality constraints on all four latent means to eliminate the statistically significant difference between Invariance Models 4 and 6 (Table 3). When standardized, the latent means in the Canadian sample range from .38 for PS to .53 for PO and represent small to medium effects in terms of Cohen's *d* (Cohen, 1988). Cohens' *d* is a commonly used measure of experimental effect and is calculated from the difference between sample means divided by the geometric mean of the sample standard deviations (Cohen, 1988; Wilkinson et al., 1999). When reliable indicators are used to generate latent means, effects slightly overestimate Cohen's metric for observed means (Hancock, 2001).

*Latent means across the age bands.* Although details of invariance testing are not shown, stratification of the respective samples into five age bands replicated the

**Table 4**  
**Structural Parameter Estimates (PE) With Respective Standard Errors (SE) for the Invariant**  
**Measurement Model (Invariance Model 4 in Table 3) Derived From the Representative 12-Subtest Samples**

	Latent Variables											
	VC			PO			WM			PS		
	PE	SE		PE	SE		PE	SE		PE	SE	
Variances and covariances of latent variables												
Verbal Comprehension												
U.S.	1 <sup>a</sup>											
Canadian	.966	.054										
Perceptual Organization												
U.S.	.613	.015	1 <sup>a</sup>									
Canadian	.549	.037	.845	.050								
Working Memory												
U.S.	.646	.022	.774	.017	1 <sup>a</sup>							
Canadian	.500	.047	.652	.047	.945	.076						
Processing Speed												
U.S.	.434	.019	.881	.008	.717	.018	1 <sup>a</sup>					
Canadian	.381	.034	.683	.041	.533	.044	.884	.051				
			Means of latent variables									
U.S.	0 <sup>a</sup>		0 <sup>a</sup>		0 <sup>a</sup>		0 <sup>a</sup>		0 <sup>a</sup>		0 <sup>a</sup>	
Canadian	.505	.038	.491	.038	.474	.046	.360	.038				

Note: VC = Verbal Comprehension; PO = Perceptual Organization; WM = Working Memory; PS = Processing Speed. In the top section of the table are shown parameter estimates for latent variable variances on the diagonal, scaled in the unit variance of the U.S. sample that was fixed for model identification. Below the diagonal are shown the covariances between latent variables. In the lower part of the table are shown the latent variable means fixed to zero in the U.S. sample. All data are derived from the U.S. and Canadian standardization samples of the Wechsler Adult Intelligence Scale—Third Edition. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Used with Permission. All rights reserved.

a. Parameter values fixed for model identification.

**Table 5**  
**Latent Means in the Canadian Sample Compared to Means**  
**in the U.S. Sample Across the Five Standardization Age Groups**

	Means of Latent Variables							
	VC		PO		WM		PS	
	PE	SE	PE	SE	PE	SE	PE	SE
<20	.52 <sup>a</sup>	(.10)	.40 <sup>a</sup>	(.10)	.45 <sup>a</sup>	(.12)	.23 <sup>a</sup>	(.10)
20-34	.67 <sup>a</sup>	(.07)	.44 <sup>a</sup>	(.07)	.39 <sup>a</sup>	(.08)	.32 <sup>a</sup>	(.07)
35-54	.49 <sup>a</sup>	(.08)	.49 <sup>a</sup>	(.08)	.41 <sup>a</sup>	(.09)	.21 <sup>a</sup>	(.08)
55-74	.31 <sup>a</sup>	(.08)	.35 <sup>a</sup>	(.09)	.36 <sup>a</sup>	(.09)	.14	(.09)
>74	.20	(.13)	.16	(.12)	.33 <sup>a</sup>	(.15)	.03	(.11)

Note: VC = Verbal Comprehension; PO = Perceptual Organization; WM = Working Memory; PS = Processing Speed. For each age group, the means in the U.S. sample that were fixed to zero (not shown). Parameter estimates (PE) with respective standard errors (SE) are shown in covariance matrix metric for the invariant measurement model (Invariance Model 4 in Table 3) estimated separately in each age stratum. All data are derived from the U.S. and Canadian standardization samples of the Wechsler Adult Intelligence Scale—Third Edition. Copyright © 1997, 2003 by the Psychological Corporation, a Harcourt Assessment company. Used with Permission. All rights reserved.

a. Parameter values different in statistically significant terms from the zero values in the U.S. sample.

same pattern of invariance to that reported above. The age bands and respective sample sizes were as follows: less than 20 (U.S.  $n = 354$ ; Canadian  $n = 157$ ), 20 to 34 (U.S.  $n = 617$ ; Canadian  $n = 360$ ), 35 to 54 (U.S.  $n = 419$ ; Canadian  $n = 252$ ), 55 to 74 (U.S.  $n = 593$ ; Canadian  $n = 216$ ), and greater than 74 (U.S.  $n = 464$ ; Canadian  $n = 107$ ). Latent variable means were then calculated from Invariance Model 4 in Table 3 and were rerun in each of the age bands; these are shown in covariance matrix metric in Table 5. Differences in the latent means were larger in the lower age groups.

*Influence of demographic variables on latent means.* To examine some of the effects of demographic differences between national samples, latent variables in the fully invariant 12-subtest measurement model (Invariance Model 4 in Table 3) were regressed on age in years, gender, and the five educational categories available for both samples (see Table 1 or the respective technical manuals: The Psychological Corporation, 2001, 2002). Gender and education category were dummy coded for this analysis. Although all the regression coefficients relating latent variables to demographic variables were statistically significant and in the expected directions, the pattern of latent mean differences remained essentially unchanged. In other words, the higher latent mean values in the Canadian sample do not appear to be a function of differences in age, gender, or education between the U.S. and

Canadian samples. In view of the differences in ethnic composition and associated coding in the two national samples, no useful comparisons could be made in relation to ethnicity.

## Discussion

Results of this study suggest that the same latent variable model underlies WAIS-III scores in the U.S. and Canadian normative samples. Consistent with previous studies using the U.S. WAIS-III standardization data (Bowden et al., 2006; Tulsy & Price, 2003) and WAIS-R data from healthy community and clinical samples (Bowden et al., 2004), a four-factor model with cross loadings for two subtests provide the best fit in the Canadian WAIS-III standardization sample. Initially the 13-subtest version was examined (see Figure 1) and then the 12 subtest version in the larger, representative samples.

Examination of the numerical components of the measurement model through incremental equality restrictions on the factor loading and observed-score intercept matrices revealed that the regression relationships relating observed scores and latent variables is the same in both versions of the test in both samples (see Tables 2 and 3). This set of findings suggests that the latent variables are estimated in equivalent scales in the U.S. and Canada. As noted above, the assumption of equivalent measurement of latent variables across samples is fundamental to generalization of validity studies and interpretation of patterns of deficit (Horn & McArdle, 1992; Meredith, 1993; Widaman & Reise, 1997). Without equivalent scales in different populations, diagnostic inferences based on the observed scores cannot be assumed to be reflecting equivalent levels of the underlying latent variable.

Examination of the invariance of the residual variances also supported retention of invariance (Invariance Model 4 in Tables 2 and 3). In the context of equivalent latent variable variances, this finding facilitates the generalization of construct validity because as is well known, variations in the reliability of test scores across groups can attenuate or otherwise distort validity relationships (Schmidt & Hunter, 1996).

Having established metric invariance of the full measurement model, with both the 12- and 13-subtest versions of the test, the values of the latent variable means, variances, and covariances were examined in the fully representative samples (Table 4). Although the overall test of equivalence of latent variances and covariances did not suggest marked differences in these structural parameters, inspection of modification indices suggested that there were statistically significant differences in some covariances, specifically VC with PO, VC with PS, and PO with WM, respectively. Because the samples can be assumed to be representative, these differences may reflect small differences in the strength of population covariances, suggesting a lesser convergence between the respective pairs of latent variables in

the Canadian population. In addition, the differences in the covariances suggest that some caution should be exercised in generalizing precise values of these specific validity correlations across populations. Studies of these relationships in Canadian clinical samples should take into account the slightly lower correlations between these latent variables in the Canadian normative sample.

Finally, latent variable means were contrasted across the fully representative samples, with the result that all differences between means were found to be statistically significant in favor of the Canadian sample, although the magnitude of these differences represented small to medium effects (Table 4). When these mean differences were reexamined in the age strata (Table 5), it was noted that more statistically significant differences between the national samples were observed in the younger age strata, with smaller differences in the older age groups. When potential differences in the measured demographic variables of age, gender, and education were incorporated into the mean-structure analysis, the pattern of mean differences remained essentially unchanged. These findings suggest that in representative samples of U.S. and Canadian participants, scores on the four cognitive variables are higher in Canada, with larger differences being observed in the younger age groups.

Iverson, Lange, and Viljoen (2006) examined the interpretive effects of applying U.S. and Canadian normative systems in a sample of 100 Canadian forensic psychiatry and neuropsychiatry inpatients. Overall, the Canadian normative system yielded scores that were systematically lower than the U.S. scores. In other words, a raw score interpreted as average (50th percentile) against the U.S. norms would be interpreted as below average (<50th percentile) against the Canadian norms. Thus, clinicians will infer slightly lower intellectual abilities from the same raw scores when using the Canadian norms. On the basis of the information available to Iverson and colleagues, the differences in the normative scores might be interpreted as due to real differences between the two normative samples, differences attributable to the different statistical norming methods, or both.

However, the results of the present study were based on raw scores derived from representative U.S. and Canadian samples and are not confounded by any differences in the procedures by which normative scores were calculated. In the context of measurement invariance, the differences in latent variable means may reflect real, albeit small, differences in the underlying abilities in the two populations.

In conclusion, the measurement model underlying scores on the WAIS-III was shown to be invariant across the U.S. and Canadian standardization samples. This finding has important implications for interpretation of construct validity in the broadest sense and shows that validity studies conducted in one population may generalize to the other. In addition, in the context of similar latent variable variances across samples, three of the six latent variable covariances were slightly smaller in the Canadian sample. In contrast, all four latent means showed elevations in the Canadian sample, these elevations being more pronounced in the younger age groups.

## References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bowden, S. C., Cook, M. J., Bardenhagen, F. J., Shores, E. A., & Carstairs, J. R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence, 33*, 363-389.
- Bowden, S. C., Weiss, L. G., Holdnack, J. A., & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adults Intelligence Scale-III. *Psychological Assessment, 18*, 334-339.
- Byrne, B. M. (1998). *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic.
- Dudgeon, P. L. (2004). A note on extending Steiger's (1998) multiple sample RMSEA adjustment to other noncentrality parameter-based statistics. *Structural Equation Modelling, 11*, 305-319.
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373-388.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.
- Iverson, G. L., Lange, R. T., & Viljoen, H. (2006). Comparing the Canadian and American WAIS-III normative systems in inpatient neuropsychiatry and forensic psychiatry. *Revue canadienne des Sciences du comportement, 38*, 348-353.
- Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 373-402). New York: Guilford.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.
- Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (Version 3). Los Angeles: Author.
- The Psychological Corporation. (2001). *WAIS-III Canadian technical manual*. San Antonio, TX: Author.
- The Psychological Corporation. (2002). *WAIS-III WMS-III technical manual* (updated edition). San Antonio, TX: Author.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.
- Saklofske, D. H., Hildebrand, D. K., & Gorsuch, R. L. (2000). Replication of the factor structure of the Wechsler Adult Intelligence Scale – Third Edition with a Canadian sample. *Psychological Assessment, 12*, 436-439.
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.
- Taub, G. E., McGrew, K. S., & Witta, E. L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-Third Edition. *Psychological Assessment, 16*, 85-89.
- Tulsky, D. S., & Ledbetter, M. F. (2000). Updating to the WAIS-III and WMS-III: Considerations for research and clinical practice. *Psychological Assessment, 12*, 253-262.

- Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six-factor model of cognitive functioning. *Psychological Assessment, 15*, 149-162.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurements invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advance from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.
- Wilkinson, L., & the Task Force on Statistical Inference: APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.