Journal of the International Neuropsychological Society (2009), **15**, 416–425.

doi:10.1017/S1355617709090602

Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery

HOLLY A. TUOKKO,[1,2] PAK HEI BENIDITO CHOU,[2] STEPHEN C. BOWDEN,[3] MARTINE SIMARD,[4] BERNADETTE SKA,[5] AND MARGARET CROSSLEY[6]

[1]Department of Psychology, University of Victoria, Victoria, British Columbia, Canada
[2]Centre on Aging, University of Victoria, Victoria, British Columbia, Canada
[3]Department of Psychology, School of Behavioural Science, University of Melbourne, Parkville, Victoria, Australia
[4]École de Psychologie, Université Laval, Québec City, Québec, Canada
[5]Faculté de Médecine—École d'orthophonie et audiologie, Université de Montréal, Montréal, Québec, Canada
[6]Department of Psychology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

(RECEIVED June 17, 2008; FINAL REVISION January 16, 2009; ACCEPTED January 16, 2009)

Abstract

Neuropsychological batteries are often translated for use across populations differing in preferred language. Yet, equivalence in construct measurement across groups cannot be assumed. To address this issue, we examined data from the Canadian Study of Health and Aging, a large study of older adults. We tested the hypothesis that the latent variables underlying the neuropsychological battery administered in French or English were the same (invariant). The best-fitting baseline model, established in the *English-speaking Exploratory sample* ($n = 716$), replicated well in the *English-speaking Validation sample* ($n = 715$), and the *French-speaking sample* (FS, $n = 446$). Across the *English-* and FSs, two of the factors, Long-term Retrieval and Visuospatial speed, displayed invariance, that is, reflected the same constructs measured in the same scales. In contrast, the Verbal Ability factor showed only partial invariance, reflecting differences in the relative difficulty of some tests of language functions. This empirical demonstration of partial measurement invariance lends support to the continued use of these translated measures in clinical and research contexts and illustrates a framework for detailed evaluation of the generality of models of cognition and psychopathology, across groups of any sort. (*JINS*, 2009, *15*, 416–425.)

Keywords: Psychometric, Elderly, Languages, Cognition, Aged, Cross-cultural comparisons

INTRODUCTION

Many clinical neuropsychological batteries used in evaluating age-associated cognitive disorders such as Alzheimer's disease were developed only in English. It has been common practice for batteries to be translated for use across populations differing in preferred language. This practice assumes that the translated battery measures the same constructs and the test scores will have the same meaning [e.g., American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME), 1999; Ardila et al., 2002].

Correspondence and reprint requests to: Holly A. Tuokko, Centre on Aging, Sedgewick Building, Room A104, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia, Canada V8W 2Y2. E-mail: htuokko@uvic.ca

Not surprisingly, use of translated tests has met with criticism. In translating test instruments, cultural differences are not necessarily taken into consideration. Similarly, the structure of the language and even minor aspects of administration or response, such as the amount of energy or time required to speak the words, may influence performance on tests. Therefore, it cannot be assumed that direct translation produces a version of a test that is equivalent in content, difficulty level, or reliability and validity to the original (AERA, APA, NCME, 1999).

It has been well documented that neuropsychological test scores from ethnic minorities can result in over- or under-identification of disorders of cognition. Bravo and Hébert (1997), using data from the Canadian Study of Health and Aging (CSHA), observed that participants assessed in French performed differently than participants assessed in English

on the Modified Mini-Mental State (3MS) Examination. Similarly, Tuokko et al. (1995) found that the rates of dementia diagnosis differed between English- and French-speaking participants in the CSHA. At that time, it was unclear whether these observed differences reflected true differences or whether they merely reflected language bias in the measures used for evaluating cognitive disorders.

The question remains whether tests, administered in different languages, can be viewed and interpreted confidently as reflecting similar underlying constructs. One necessary, but not sufficient, condition for determining test fairness, in this situation, is measurement invariance (Meredith & Teresi, 2006). Measurement invariance, or equivalence, refers to the extent to which test items are perceived and interpreted similarly by test takers in different groups and reflect the same underlying psychological constructs across groups (Byrne & Watkins, 2003). The term "measurement equivalence" is used to convey the demonstration of equivalent construct measurement in a set of test scores across groups. However, the process of establishing measurement equivalence involves several detailed steps in testing to reject the null hypothesis of measurement invariance (Byrne, 1998). Each step of invariance testing involves explicit evaluation of the similarity of construct measurement, using techniques that are not well known among applied and clinical researchers (Bontempo & Hofer, 2007; Bowden et al., 2004). For reasons of technical precision, we use the term invariance in the remainder of this article, but readers can assume that if the null hypothesis of invariance is retained, then the inference of equivalent construct measurement is justified.

Establishing measurement invariance is essential if one wishes to make meaningful group comparisons. The extent to which a measure is not invariant across groups limits any interpretation of between-group differences. Establishing measurement invariance is a precursor to the investigation of classification or diagnostic validity (Meredith & Teresi, 2006; Vandenberg & Lance, 2000). Given its importance, it is no longer acceptable to presume measurement invariance; rather, measurement invariance needs to be demonstrated and established. This issue is relevant to all application of psychological tests across groups that may differ on the measured constructs, has obvious relevance to neuropsychological assessment, and to assessment across languages or cultures (Bontempo & Hofer, 2007; Bowden et al., 2008a).

The multigroup confirmatory factor analysis (CFA) framework is one method to test whether a measure is invariant across groups (Teresi, 2006). Following the work of Meredith (1993) and Widaman and Reise (1997), assessment of invariance entails four increasingly restrictive tests: configural and then weak, strong (or scalar), and strict invariance. This hierarchy of tests provides increasing evidence of measurement invariance and, more importantly, determines the types of group comparisons that are defensible (Chen et al., 2005; Meredith & Teresi, 2006; Widaman & Reise, 1997). For applications in clinical neuropsychology, see Bowden et al. (2004), Bowden et al. (2008a), Gladsjo et al. (2004), Meredith and Teresi (2006), Reilly et al. (2006), and Widaman and Reise (1997).

Brown (2006) provided an expository account of the steps involved, the similarities and differences between exploratory and CFAs, and how invariance analysis is an extension of these methods. Within the factor analytic framework, configural invariance, the most basic level, can be tested by specifying the same number of factors and the same assignment of test scores (or indicators) to factors in each group; however, the factor loadings can differ across group. Configural invariance implies that different groups have similar but not identical latent structure. The second level, weak invariance, requires that the numerical values of factor loadings are identical across group and implies that comparisons across groups in terms of correlation-based, criterion-related validity are permissible. Next, strong invariance is said to apply if, in addition, the values of the observed variables intercepts are equivalent; construct validity in its broadest sense can be evaluated with the knowledge that the same constructs are measured in both groups and the respective constructs are measured on the same numerical scale across groups. In other words, strong invariance is necessary for valid interpretation of mean differences, for example, deficits in clinical groups, and for evaluation of convergent and discriminant validity (Meredith & Teresi, 2006; Widaman & Reise, 1997). Finally, strict invariance is said to hold when residual variances are equivalent across groups. When strict invariance is met, any observed differences between groups are attributed only to group differences at the latent variable level but not their different residual variances including different reliabilities.

The purpose of this study was to determine whether the latent variable structure or construct measurement of a neuropsychological battery, administered in English or French, was invariant. We report a secondary analysis of data derived from the CSHA, a large, longitudinal, epidemiological study of cognitive impairment and dementia. This data set provides a unique opportunity to examine measurement invariance of a neuropsychological battery translated and administered in different languages within a nationwide study. Demonstration of measurement invariance in this context will inform the interpretation and use of translated measures in clinical practice and serves as a model for addressing the empirical determination of invariance that could be applied in other studies where linguistic or cultural differences are present.

When identifying the latent variable model that underlies a selection of tests chosen for clinical purposes, there will be a trade-off between construct representation (Whitely, 1983) and pragmatic concerns. When testing older participants, time constraints and patient tolerance are paramount. Thus, it may be impractical to administer a set of tests that provide a detailed elaboration of a well-known model of ability such as the Cattell–Horn–Carroll (C-H-C) model (Carroll, 1993; Flanagan & Harrison, 2005). In other words, it may be necessary to accept a baseline model that does not clearly articulate the five or six latent variables underlying a detailed ability battery (Bowden et al., 2004; Carroll, 1993; Flanagan & Harrison, 2005).

MATERIALS AND METHODS

Sample

This study used data collected as part of the first wave of the CSHA (Canadian Study of Health and Aging Working Group, 1994). Briefly, the first wave of the CSHA was an epidemiological study concerned with the prevalence of dementia in Canadians aged 65 years and older. In 1991, a stratified random sample of 9008 community-dwelling older adults and 1255 institutional residents took part in this study. Community participants were administered the 3MS (Teng & Chui, 1987) to screen for cognitive impairment. Those community participants who scored below 78 on the 3MS, a random sample of 494 who scored 78 or above, and all institutional residents were clinically examined (n = 2914). The clinical examination was designed to confirm the diagnosis of cognitive impairment and to provide a differential diagnosis of dementia. The assessment protocol included a clinical neurological examination, nurse's evaluation, blood tests, and a neuropsychological assessment. The neuropsychological assessment was administered to those participants who scored more than 50 on the 3MS (N = 1590; 50 < 3MS < 78, n = 1096; 3MS ≥ 78, n = 494). The Canadian Tri-Council Policy Statement on the Ethical Conduct for Research Involving Humans was followed in the conduct of this research.

Study Variables

The CSHA neuropsychological battery was selected by a group of neuropsychologists from across Canada to reflect the *Diagnostic and Statistical Manual for Mental Disorders*, third edition, revised (American Psychiatric Association, 1987), criteria for dementia. The final selection of tests consisted of 11 measures that assessed memory, abstract thinking, judgment, language, and construction (Tuokko & Woodward, 1996; Tuokko et al., 1995). Memory was assessed using the Information subtest of the Wechsler Memory Scale (Wechsler, 1945), the Benton Visual Retention Test-Multiple Choice version (Benton, 1974), a modified Buschke's Free Recall total score (Buschke, 1984; Tuokko et al., 1991), and the Rey Auditory Verbal Learning Test total (Rey AVLT; Bleecker & Bolla-Wilson, 1988; Rey, 1964). Short forms of three Wechsler Adult Intelligence Scale–Revised (WAIS-R; Wechsler, 1981) subtests were included to assess abstract thinking, judgment, and construction (Block Design, Comprehension, and Similarities; Satz & Mogel, 1962), and the WAIS-R Digit Symbol subtest was included as a measure of psychomotor speed. Measures of language abilities included Animal Naming (Rosen, 1980), Controlled Oral Word Association Test (COWAT; Spreen & Benton, 1977), and the Token Test (Benton & Hamsher, 1989). All measures were either available in French or translated from English to French.

Data Analysis

Since the factor structure of the CSHA battery has not been examined before, we first identified and evaluated candidate factor models derived from exploratory factor analysis (EFA) and theory. Establishing a baseline model with reasonable fit to the data is the first step in conducting invariance analyses (Byrne, 1998; Brown, 2006). The baseline model served as a point of comparison for departure in model fit when we introduced constraints on different aspects of the model to test for measurement invariance. For this study, the entire CSHA English-speaking sample, including cases with missing data on any subset of variables (n = 1431), was randomly divided into halves (using SPSS 14.0; SPSS Inc., 2005): (1) *English-speaking Exploratory sample* (EES) and (2) *English-speaking Validation sample* (EVS). These samples were used to explore and replicate, respectively, a baseline model of the 11 CSHA battery scores included in this analysis. Replication of factor analysis is a critical and often neglected step in establishing a generalizable latent structure (Bowden et al., 2004; Bowden et al., 2008b; Brown, 2006; Preacher & McCallum, 2003; Strauss & Smith, in press; Vandenberg & Lance, 2000). Having established a replicable latent variable structure in the randomly chosen English-speaking subsamples, the invariance of the model was then examined in the *French-speaking sample* (FS).

Although we could have chosen to combine the English subsamples to compare with the FS, use of unequal sample sizes in multiple-group CFA can lead to bias in parameter estimates, favoring the large sample. For this reason, it is preferable to use samples of approximately equal size (Brown, 2006). Therefore, once the best-fitting model was selected, we fitted this model to both the EES and the FS simultaneously without any equality constraints placed across groups. This unconstrained model became the baseline model. As explained below, we introduced increasingly restrictive equality constraints on the factor loadings, followed by the intercepts, and finally on the residual variances and covariance between the two samples. Whether or not the battery was invariant, and if so, at what level, was evaluated by the reduction in model fit between these restricted models and the baseline model.

The EES comprised 716 participants, the EVS comprised 715 participants, and the FS comprised 446 participants. One study participant was not included in the analysis because this participant was missing all cognitive scores. In the remaining participants, missing data were treated as missing completely at random. Basic demographic characteristics of the samples are given in Table 1. Based on available data, the FS was slightly younger than their English counterparts, $F(2,1874)$ = 10.02, $p < .001$. In addition, participants whose preferred language of testing was French had fewer years of education compared to those who preferred English, $F(2,1272) = 48.57$, $p < .001$. Finally, English-speaking participants were more likely to be living alone, $\chi^2(2, N = 1295) = 27.10$, $p < .001$.

RESULTS

Baseline Model Identification

The EFA computed for the EES used maximum likelihood estimation and an oblique, rotated solution provided by default

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

Table 1. Demographic characteristics of the study sample

Characteristics	EES (n = 716)	EVS (n = 715)	FS (n = 446)
Age, mean (*SD*)	80.3 (7.3)	80.5 (7.0)	78.7 (6.6)
Education, mean (*SD*)	9.3 (4.0)	9.5 (4.1)	6.9 (3.6)
Sex, count (%)			
Female	429 (59.9)	449 (62.8)	297 (66.6)
Live alone, count (%)			
Yes	223 (44.2)	217 (44.7)	85 (27.8)
Maternal language, count (%)			
English	559 (78.4)	557 (78.3)	1 (0.2)
French	34 (4.8)	32 (4.5)	424 (96.1)

Note. The total percentage for maternal language does not add up to 100 due to report of bilingualism and other languages.

with Mplus 5.1 (Muthen & Muthen, 1998–2007). Models derived from EFA were then converted to simple-structure CFA models (Brown, 2006) and compared with *a priori* CFA models derived from previous research. Relative model fit of CFAs was examined in terms of the two-index strategy of Hu and Bentler (1998), although this strategy may be conservative (Marsh et al., 2004). This strategy places most emphasis on the standardized root mean square residual (SRMR) or root mean square error of approximation (RMSEA). Other goodness-of-fit statistics reported for model evaluation included the comparative fit index (CFI), the Tucker–Lewis Index or non-normed fit index (TLI or NNFI), and the expected cross-validation index (ECVI; Browne & Cudeck, 1993).

For the invariance analyses, we also examined the multiple-group versions of the RMSEA, Gamma 1 (a multiple-group version of the goodness-of-fit index), and the ECVI (Dudgeon, 2004). Since small sample statistics such as chi-square are overly sensitive to changes in goodness-of-fit when applied to large samples (Cheung & Rensvold, 2002), most attention was directed to the overall pattern of fit (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000) and indices for which confidence intervals can be estimated. 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 the fit statistics were estimated using Mplus 5.1 (Muthen & Muthen, 1998–2007), except for the ECVI, and confidence intervals for the RMSEA, ECVI, and Gamma 1, which were calculated using a program provided by Dudgeon (2004).

From the EFA, a five-factor model was found to fit data from the EES, $\chi^2(10, n = 716) = 15.44$, $p = .117$. However, the corresponding EFA as CFA model would not converge. This may have occurred because the five-factor EFA model had only one indicator for each of two factors and was not identified (Brown, 2006). Nevertheless, to ensure that a viable model was not overlooked, the five-factor EFA was then specified as a simple-structure CFA (Table 2) and the standardized residual variances for the two tests loading on the single-item factors were fixed to .3, respectively, to identify the model. In this and all subsequent simple-structure CFAs, two correlated residual terms were freed for estimation to account for method variance in items with similar administration or response formats (Bowden et al., 1999; Brown, 2006). The correlated residuals were between Animal Naming and COWAT and between the Buschke Free Recall total and the Rey AVLT total, respectively. Again the simple-structure CFA for this five-factor model did not provide an admissible solution, being unable to estimate the correlation between the separate fluency factors. Therefore, the five-factor model was not considered further.

The four-factor simple-structure CFA model corresponding to the four-factor EFA model from the EES was just identified with at least two indicators per factor (Table 2). However, this model had a correlation between the Visuospatial speed and Fluency factors less than 2 *SE*s from unity and was therefore considered unacceptable. The three-factor simple-structure CFA, corresponding to the EFA result, also had one factor with only one indicator (Table 3), the Rey AVLT score, the standardized residual for which was fixed at .3 for identification. Fit statistics for this model are shown in Table 3.

An *a priori* three-factor CFA model based on previous analyses of Wechsler Intelligence and Memory Scales and C-H-C theory (Table 2; also Bowden et al., 1997; Flanagan & Harrison, 2005) provided a chi-square and other fit indices that were similar in absolute terms to the chi-square and fit indices for the three-factor model from the EFA above (Table 3). However, the *a priori* model is preferred on theoretical and substantive grounds since the three-factor model from the EFA places the Rey AVLT score on a factor of its own, separate from the factor on which the conceptually and theoretically similar Buschke score loads. Such anomalous results are a recognized feature of EFA because EFA confounds reliable variance with error variance and is therefore particularly susceptible to anomalous solutions (Preacher & MacCallum, 2003; Brown, 2006).

A two-factor simple-structure CFA (Table 2) corresponding to the EFA result is identical to a two-factor model that might have been specified *a priori* (Bowden et al., 1997; Flanagan & Harrison, 2005) and was significantly poorer fitting than the nested three-factor model (Table 3). A single-factor model was poorer fitting than the two-factor model (Table 3).

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

Table 2. Alternative models examined in the EES

Test score	Two-factor simple-structure CFA from EFA	Three-factor *a priori* CFA	Three-factor simple-structure CFA from EFA	Four-factor simple-structure CFA from EFA	Five-factor simple-structure CFA from EFA
WMS-R Information	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval
Rey AVLT total recall	Long-term Retrieval	Long-term Retrieval	Rey AVLT memory	Long-term Retrieval	Long-term Retrieval
Buschke total recall	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval	Long-term Retrieval
BVRT correct	General ability	Visuospatial speed	General ability	Visuospatial speed	Visuospatial speed
WAIS-R Comprehension	General ability	Verbal Ability	General ability	Verbal Ability	Verbal Ability
WAIS-R Similarities	General ability	Verbal Ability	General ability	Verbal Ability	Verbal Ability
WAIS-R Block Design	General ability	Visuospatial speed	General ability	Visuospatial speed	Visuospatial speed
WAIS-R Digit Symbol	General ability	Visuospatial speed	General ability	Visuospatial speed	Visuospatial speed
Token Test	General ability	Verbal Ability	General ability	Verbal Ability	Verbal Ability
COWAT	General ability	Verbal Ability	General ability	Fluency	Verbal Fluency
Animal Naming	General ability	Verbal Ability	General ability	Fluency	Animal Fluency

The above-mentioned analysis of simple-structure and *a priori* CFA models was repeated for the EVS. A similar pattern of results in terms of incremental model fit was observed to that reported for the EES and led to selection of the *a priori* three-factor model as best fitting in both English-speaking subsamples (Model 3a in Table 3). The analyses were then rerun in the FS, with a similar pattern of relative fit between models (Table 3).

Prior to examining measurement invariance, Model 3a was examined for potential *post hoc* improvements. Only *post hoc* modifications that were theoretically meaningful and significant in all three samples were regarded as acceptable with the caveat that across linguistic groups, there may be differences in method variance (Byrne, 1998). In addition to the *a priori* correlated residuals described above, three additional correlated residuals were identified for inclusion in a *post hoc* model. These included correlated residuals between Buschke total and the WMS Information Score, which, like the *a priori* correlated residuals, may be interpreted as method variance in long-term memory tests. Parameters were also included for the correlated residuals between COWAT and the Rey AVLT and between Animal Naming and Buschke,

Table 3. Goodness-of-fit statistics for the baseline model estimation for the 11 neuropsychological tests in the Canadian Study on Health and Aging Wave 1 data in the EES (n = 716), EVS (n = 715), and FS (n = 446)

	χ^2	*df*	CFI	TLI	lower bounds (95%) (RMSEA) upper bounds (95%)	SRMR	lower bounds (95%) (ECVI) upper bounds (95%)
1. Single-factor CFA							
EES	469.62*	42	.879	.841	.1079 (.1193) .1310	.063	.6148 (.7251) .8492
EVS	484.38*	42	.874	.835	.1100 (.1215) .1331	.062	.6345 (.7468) .8731
FS	234.46*	42	.891	.858	.0866 (.1015) .1168	.056	.5187 (.6377) .7795
2. Two-factor CFA from EFA							
EES	422.20*	41	.892	.855	.1024 (.1140) .1259	.058	.5578 (.6616) .7795
EVS	436.29*	41	.887	.849	.1046 (.1162) .1281	.057	.5763 (.6823) .8021
FS	222.07*	41	.898	.863	.0845 (.0996) .1151	.055	.4992 (.6145) .7527
3a. Three-factor *a priori* CFA							
EES	328.42*	39	.918	.884	.0899 (.1019) .1141	.055	.4459 (.5361) .6403
EVS	297.60*	39	.926	.896	.0843 (.0964) .1087	.049	.4083 (.4937) .5931
FS	178.64*	39	.921	.889	.0740 (.0897) .1058	.052	.4250 (.5261) .6503
3b. Three-factor *post hoc* CFA							
EES	134.11*	35	.972	.956	.0496 (.0629) .0765	.030	.2231 (.2758) .3426
EVS	151.41*	35	.967	.948	.0551 (.0683) .0817	.030	.2433 (.3004) .3716
FS	73.64*	35	.978	.966	.0303 (.0498) .0686	.030	.2537 (.3087) .3874
4. Three-factor CFA from EFA							
EES	339.66*	40	.915	.883	.0905 (.1024) .1145	.053	.4570 (.5490) .6549
EVS	315.93*	40	.921	.892	.0864 (.0983) .1104	.048	.4283 (.5166) .6187
FS	190.96*	40	.915	.883	.0767 (.0921) .1079	.052	.4439 (.5492) .6774

Notes. Models compared in this table comprise simple-structure CFAs derived from EFA in the EES or CFA models derived from previous research. Three samples were examined, an EES (n = 716), an EVS (n = 715), and the FS (n = 446).
*p < .01 for chi-square test.

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

respectively, likely reflecting retrieval processes from long-term memory. Finally, model fit was significantly improved in all samples by allowing Animal Naming to load jointly on Long-term Retrieval and the Verbal Ability factors. The final model fit information for the three-factor model with *post hoc* modifications (labeled Model 3b) is shown in Table 3 and indicates a good fit in all three samples.

Invariance Analysis

EES versus *FS*

Following the strategy outlined by Widaman and Reise (1997), we next examined measurement invariance (Bowden et al., 2004; Bowden et al., 2008b). The baseline model was estimated first (shown as Invariance Model 1, Table 4), the reported chi-square statistics reflecting the sum of chi-square statistics and degrees of freedom observed in the separate samples (Model 3b in Table 3). The measurement model was reestimated holding all elements of the factor loading, raw score intercept, and residual variance matrices to equality across groups. Imposing these equality constraints on all elements of the measurement model, including residual variances and covariances, resulted in significant loss of fit compared to the baseline invariance model (Invariance Model 1 in Table 4), difference $\chi^2(43, N = 1162) = 258.66$, $p < .05$.

To determine which elements of the measurement model led to this significant loss of fit, the model was reestimated in separate steps, first holding only the matrix of factor loadings to equality across groups (Invariance Model 2 in Table 4). Examination of fit with factor loadings held invariant revealed significant loss of fit, difference $\chi^2(9, N = 1162) = 26.67$, $p < .01$, compared to the baseline model (Invariance Model 1 in Table 4). In contrast, the other fit indices indicate no appreciable loss of fit. In particular, there was no change in the value of CFI or TLI, and the fit indices (where confidence intervals available) showed no significant change. These results suggested retention of the assumption of weak invariance across groups.

Next, examination of fit after additional equality constraints were imposed on the intercepts (Invariance Model 3a in Table 4) indicated a highly significant loss of fit, difference $\chi^2(8, N = 1162) = 130.68$, $p < .01$, compared to the preceding step (Invariance Model 2). In addition, there was an appreciable change in the CFI in terms of the criteria outlined above, and each of the fit indices, where confidence intervals could be calculated, showed significant change from Invariance Model 2. Examination of modification indices for Invariance Model 3a indicated that two of the intercepts for the Verbal Ability factor were not invariant. In order of largest modification index, freeing the intercept for WAIS-R Similarities, difference $\chi^2(1, N = 1162) = 75.87$, $p < .01$, and then for Token Test, difference $\chi^2(1, N = 1162) = 32.29$, $p < .01$, resulted in a model with partially invariant intercepts (Invariance Model 3b in Table 4). In terms of the criteria outlined above, Invariance Model 3b shows no appreciable loss of fit compared to Invariance Model 2 (Table 4).

Finally, additional equality constraints were imposed on Invariance Model 3b to constrain all residual variances and covariances to equality across groups (Invariance Model 4a in Table 4). Although the residual covariances are commonly tested separately because these parameters may reflect sample-specific responses, the residual covariances incorporated in this analysis may instead be interpreted as reflections of method variance (Brown, 2006; Cortina, 2002). That is, the correlated residuals in the model reflect variance related to test administration and response characteristics (*viz.*, fluency tests and list learning, respectively) or retrieval from long-term memory (Carroll, 1993) rather than unique sample characteristics. Therefore, it is a stronger test of the measurement model to demonstrate invariance of residual variances and covariances, and so, for simplicity, the test of invariance

Table 4. Summary of tests for metric invariance of ability measurement in 11 raw scores across the CSHA EES ($n = 716$) and the CSHA FS ($n = 446$)

Invariance model	χ^2	*df*	CFI	TLI	lower bounds (95%) (RMSEA) upper bounds (95%)	SRMR	lower bounds (95%) (ECVI) upper bounds (95%)	lower bounds (95%) (Gamma 1) upper bounds (95%)
Model 1—Baseline (configural) Invariance	207.75	70	.974	.959	.0474 (.0582) .0692	.030	.2865 (.3269) .3761	.9704 (.9789) .9859
Model 2—Invariant factor loadings	234.42	79	.971	.959	.0480 (.0582) .0685	.043	.2908 (.3340) .3861	.9674 (.9762) .9837
Model 3a—Model 2 and invariant intercepts	366.00	87	.948	.934	.0651 (.0744) .0838	.061	. 3765 (.4334) .4991	.9474 (.9581) .9676
Model 3b—Model 2 and partially invariant intercepts	257.85	85	.967	.958	.0494 (.0592) .0691	.044	.2980 (.3437) .3981	.9644 (.9736) .9815
Model 4a—Model 3b and invariant residual variances and covariances	358.51	101	.951	.947	.0575 (.0663) .0752	.075	.3468 (.4023) .4665	.9506 (.9612) .9705
Model 4b—Model 3b and partially invariant residual variances	314.23	100	.960	.956	.0518 (.0608) .0699	.064	.3149 (.3659) .4257	.9575 (.9675) .9762

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

on the residual variances and covariances is combined. The fit statistics associated with Invariance Model 4a were significantly different from Model 3b (Table 4), difference $\chi^2(16, N = 1162) = 100.66$, $p < .01$. In addition, there was appreciable change in the CFI; the ECVI and Gamma 1 both fell outside the 95% confidence interval for Invariance Model 3b. As such, there was evidence to reject the assumption of equality of residual variances or covariances. Again, modification indices were examined and indicated that freeing the residual variance for COWAT, difference $\chi^2(1, N = 1162) = 44.29$, $p < .01$, resulted in a model (Invariance Model 4b in Table 4), not appreciably different, in terms of the overall pattern of fit indices, from Invariance Model 3b.

In the context of two fully invariant latent variables, Long-term Retrieval and Visuospatial speed, and one partially invariant latent variable, Verbal Ability, invariance of latent variable variances and covariances, and latent variable means was tested. Invariance Model 4b was modified to hold all factor variances and covariances to equality, resulting in a significant loss of fit, difference $\chi^2(6, N = 1162) = 24.00$, $p < .01$. Using the modification indices as a guide, the variances of the Verbal Ability and Visuospatial speed factors and the covariance of Verbal Ability with Visuospatial speed were freed for separate estimation across groups. The resulting model was not significantly different from Invariance Model 4b, difference $\chi^2(3, N = 1162) = 6.40$, $p > .05$.

To examine the invariance of the latent means, Invariance Model 4b was modified to hold the latent means to equality across groups. This led to a significant loss of fit, difference $\chi^2(3, N = 1162) = 138.06$, $p < .01$. Modification indices suggested freeing the latent means for Verbal Ability and Visuospatial speed. This resulted in a nonsignificant difference from Invariance Model 4b, difference $\chi^2(1, N = 1162) = 0.01$, $p > .05$, indicating that the latent mean for Long-term Retrieval only could be considered equivalent across groups. When estimated from the partially invariant measurement model (Model 4b in Table 3), the latent means for Verbal Ability and Visuospatial speed, expressed in completely standardized metric, were 0.89 and 0.32 *SD* units, respectively, lower in the FS compared to the EES. These differences in latent means slightly overestimate the metric of Cohen's *d* (Bowden et al., 2008b; Hancock, 2001).

Factor loadings for the partially invariant measurement model (Invariance Model 4b in Table 4) are shown in Table 5, separately for each sample. Factor loadings are estimated separately and differ in absolute numerical terms because of numerical differences in some intercepts, residual variances, and factor variances and covariances, as detailed above. Nevertheless, as the above-mentioned invariance analysis suggests, most of the factor loadings do not differ significantly across samples, where significance is defined here as difference of more than 2 *SE*s.

DISCUSSION

The best-fitting baseline model found in our EES fitted well in the other samples. This result was a little surprising in view of the pragmatic constraints on testing cognitive abilities in older adults, some of whom were in advanced old age (Bowden et al., 1999). Underrepresentation of constructs in shorter batteries necessitates combinations of abilities that might be identified as separate abilities in a more extensive test battery (Whitely, 1983). The best-fitting model provides further evidence of the generality of a well-validated model of cognitive abilities (Carroll, 1993; Flanagan & Harrison, 2005).

Having established a model that replicated well in all samples, measurement invariance was then examined across the English-speaking sample and FS. The finding of equality of factor loadings (Invariance Model 2 in Table 4) is of practical importance because it allows comparison of convergent and discriminant validity relations across groups (Meredith, 1993; Widaman & Reise, 1997). It is these relations that are of most interest to clinicians and of critical importance to the criterion-related validity of findings arising from the CSHA study. The finding of invariance of factor loadings underlying all the cognitive abilities in both the English- and the French-speaking participants allows direct comparisons across groups in terms of criterion-related validity.

In contrast, the observed score intercepts were not invariant across *English* and *French* samples. The nonequivalent intercepts were all related to the Verbal Ability factor and reflect differences in level of difficulty across groups. The intercepts for Similarities and Token Test were higher (i.e., easier) in the FS compared to the EES. There was never any intention to make strict comparisons on group means across linguistic groups, and in view of the type of translations involved in adaptation of the tests, a lack of equality of intercepts is of little practical importance.

Perhaps more surprising is the presence of equality of intercepts for some of the Verbal Ability tests and for all of the Long-term Retrieval and Visuospatial speed factor, indicating that raw scores for the respective indicator variables and latent variable means are directly comparable across language groups. These findings stand in contrast to studies on the Mini-Mental State Examination where items varied in difficulty and discrimination across cultural and linguistic groups leading to over- or underidentification of dementia and cognitive impairments for these groups (Ramirez et al., 2006). The finding of strong measurement invariance for two of the three factors is testimony to the generality of some components of the underlying model of cognitive abilities, despite being examined across different linguistic groups.

Although the Long-Term Retrieval and Visuospatial speed factors may be less culturally or linguistically dependent than the Verbal Ability factor, this assumption need not hold across all cultures. Omura and Sugishita (2004), for instance, reported only configural invariance on the Wechsler Memory Scale–Revised in a standardized sample from Japan and the United States. However, the results of Omura and Sugishita (2004) are ambiguous because they used an analytic approach that differs from commonly recommended methods (e.g., Widaman & Reise, 1997). Invariance testing provides a powerful framework for examination of these fundamental aspects

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

Table 5. Standardized factor loadings and variance explained by the model (with *SE*s) in each of the 11 test scores, shown separately for the EES (n = 716) and FS (n = 446)

	Factor loadings			
	Verbal Ability, PE (*SE*)	Visuospatial Ability, PE (*SE*)	Long-term Retrieval, PE (*SE*)	Explained variance, R^2 (*SE*)
Comprehension				
English	.716 (.019)	0[a]	0[a]	.513 (.028)
French	.634 (.024)		0[a]	.401 (.031)
Similarities				
English	.774 (.017)	0[a]	0[a]	.599 (.026)
French	.699 (.023)	0[a]	0[a]	.488 (.032)
Token Test				
English	.673 (.021)	0[a]	0[a]	.453 (.028)
French	.588 (.027)	0[a]	0[a]	.345 (.032)
COWAT				
English	.693 (.022)	0[a]	0[a]	.481 (.030)
French	.733 (.023)	0[a]	0[a]	.537 (.034)
Animal Fluency				
English	.301 (.040)	0[a]	0[a]	.549 (.027)
French	.254 (.034)	0[a]	0[a]	.497 (.031)
Block Design				
English	0[a]	.742 (.018)	0[a]	.550 (.026)
French	0[a]	.677 (.022)	0[a]	.458 (.030)
Digit Symbol				
English	0[a]	.895 (.012)	0[a]	.801 (.021)
French	0[a]	.857 (.016)	0[a]	.735 (.028)
Benton VRT				
English	0[a]	.707 (.020)	0[a]	.500 (.028)
French	0[a]	.639 (.025)	0[a]	.408 (.032)
WMS Information				
English	0[a]	0[a]	.641 (.024)	.410 (.031)
French	0[a]	0[a]	.620 (.027)	.385 (.033)
Buschke Free Recall				
English	0[a]	0[a]	.640 (.029)	.409 (.037)
French	0[a]	0[a]	.620 (.032)	.384 (.040)
Rey AVLT				
English	0[a]	0[a]	.790 (.021)	.624 (.033)
French	0[a]	0[a]	.774 (.025)	.599 (.039)
Animal Fluency				
English	0[a]	0[a]	.503 (.040)	.549 (.027)
French	0[a]	0[a]	.503 (.040)	.497 (.031)

[a]Parameters fixed to zero.

of measurement across groups (Meredith & Teresi, 2006) and may have a broad application in clinical neuropsychology, allowing evaluation of many issues related to the generality of models of cognition and psychopathology. An alternative approach to testing invariance of loadings and intercepts is to examine each factor separately. When applied to our data, this approach led to an identical pattern of results to that reported above. However, any approach to invariance analysis that increases the number of statistical comparisons inflates the risk of Type I errors, thereby risking false rejection of the assumption of invariance (for examples of this approach, see Bowden et al., 2001; Byrne, 1998). As a consequence, the examination of each factor separately is not recommended (Vandenberg & Lance, 2000; Widaman & Reise, 1997).

The final element of the measurement model examined for invariance involved the residual variances and covariances. One of the residual parameter estimates was not invariant, but this finding may reflect sample-specific variation in the reliability of specific tests or in unique variance elements or latent variable variances (Widaman & Reise, 1997). Lack of residual invariance does not qualify the inference of strong measurement invariance, and hence does not limit generality of construct measurement when factors are derived from scalar invariant test items (Meredith, 1993; Widaman & Reise, 1997).

Therefore, it can be inferred that Long-term Retrieval and Visuospatial speed, and the Verbal Ability factor, when estimated from the scalar equivalent COWAT, Animal Fluency

https://doi.org/10.1017/S1355617709090602 Published online by Cambridge University Press

and Comprehension tests, are operating equivalently and can be viewed as reflecting the same constructs across groups. The empirical demonstration of measurement invariance lends support to the continued use of these translated measures in clinical and research contexts and to the generality of a model of cognition in diverse populations.

When latent variable variances and covariances were examined for the scalar equivalent constructs (estimated from Invariance Model 4b in Table 4), the variances of the Verbal Ability and Visuospatial speed and the covariance between Visuospatial speed and Verbal Ability were found to differ across groups. Variability of Verbal Ability and Visuospatial speed, respectively, was significantly less in the FS, whereas the covariance (or unstandardized correlation) was significantly greater. Differences across samples in the correlations between test scores are often interpreted as evidence of "dissociations," which are, in turn, interpreted as evidence of group differences in the construct composition of the test scores (Bates et al., 2003; Bowden et al., 2008a). However, unless measurement invariance is examined, such dissociations may be ambiguous. Future research may include more detailed examinations of why these differences are occurring and the impact this may have for other studies where linguistic or cultural differences are present.

Finally, the examination of latent variable means showed that Long-term Retrieval did not differ across groups. In contrast, the means for Verbal Ability and Visuospatial speed were significantly lower in the French sample (0.89 and 0.32 pooled *SD* units, respectively). The difference in the Verbal Ability mean is substantial and may reflect real differences in educational or other cultural effects associated with the achievement of crystallized verbal abilities throughout the life span of these long-lived research participants.

In conclusion, the results of this study illustrate the multiple implications for construct validity of neuropsychological assessment, which can be addressed through the examination of measurement invariance. Two important cognitive ability constructs showed strong invariance despite being administered in different languages. A third ability construct was partially invariant, some of the corresponding indicator variables reflecting differences due to translation or other cultural effects. Although still not well known among clinical researchers, the principles of measurement invariance provide a detailed framework for understanding how tests, and the underlying constructs reflected in the test scores, work across different populations.

ACKNOWLEDGMENTS

The data reported in this article were collected as part of the first wave of the CSHA. This was funded by the Seniors Independence Research Program, administered by the National Health Research and Development Program of Health and Welfare Canada (Project No. 6606-3954-MC[S]). The study was coordinated through the University of Ottawa and the federal government's Laboratory Centre for Disease Control. A research personnel award from the Canadian Institutes of Health Research, Institute of Aging, provided support for H.A.T. in the preparation of this manuscript. Additional support was provided by a Michael Smith Foundation for Health Research award for research unit infrastructure held by the Centre on Aging at the University of Victoria. Funding to support involvement of P.H.B.C. in this study was provided by the Canadian Institutes of Health Research through their support of the Canadian Longitudinal Study on Aging. No significant financial interests or other relationships exist for any of the authors, which might be interpreted as having influenced the research. This manuscript and the information in it have never been published either electronically or in print.

REFERENCES

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

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., revised.). Washington, DC: American Psychiatric Association.

Ardila, A., Rodriguez-Menéndez, G., & Rosselli, M. (2002). Current issues in neuropsychological assessment with hispanics/latinos. In F.R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (pp. 161–179). Lisse, The Netherlands: Swets & Zeitlinger.

Bates, E., Appelbaum, M., Salcedo, J., Saygin, A.P., & Pizzamiglio, L. (2003). Quantifying dissociations in neuropsychological research. *Journal of Clinical and Experimental Neuropsychology*, *25*, 1128–1153.

Benton, A.L. (1974). *Revised visual retention test: Clinical and experimental applications*. New York: Psychological Corporation.

Benton, A.L. & Hamsher, K. (1989). *Multilingual Aphasia Examination: Manual of Instructions*. Iowa City, IA: AJA Associates.

Bleecker, M.L. & Bolla-Wilson, K. (1988). Age-related sex differences in verbal memory. *Journal of Clinical Psychology*, *44*, 403–411.

Bontempo, D.E. & Hofer, S.M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A.D. Ong & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 153–175). New York, NY: Oxford University Press.

Bowden, S.C., Carstairs, J.R., & Shores, E.A. (1999). Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale–Revised and Wechsler Memory Scale–Revised scores in a healthy community sample. *Psychological Assessment*, *11*, 339–344.

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., Dodds, B., Whelan, G., Long, C.M., Dudgeon, P., Ritter, A.J., & Clifford, C.C. (1997). Confirmatory factor analysis of the Wechsler Memory Scale–Revised in a sample of alcohol dependent clients. *Journal of Clinical and Experimental Neuropsychology*, *19*, 755–762.

Bowden, S.C., Gregg, N., Bandalos, D., David, M., Coleman, C., Holdnack, J.A., & Weiss, L.G. (2008a). Latent mean and covariance differences with measurement equivalence in college students with developmental difficulties versus the Wechsler Adult Intelligence-III/Wechsler Memory Scale-III normative sample. *Educational and Psychological Measurement*, *68*, 621–642.

Bowden, S.C., Ritter, A.J., Carstairs, J.R., Shores, E.A., Pead, J., Greeley, J.D., Whelan, G., Long, C.M., & Clifford, C.C. (2001). Factorial invariance for combined WAIS-R and WMS-R scores in a sample of patients with alcohol dependency. *The Clinical Neuropsychologist*, *15*, 69–80.

Bowden, S.C., Weiss, L.G., Holdnack, J.A., Bardenhagen, F.J., & Cook, M.J. (2008b). Equivalence of a measurement model of

cognitive abilities in US standardization and Australian neuroscience samples. *Assessment*, *15*, 132–144.
Bravo, G. & Hébert, R. (1997). Reliability of the Modified Mini-Mental State Examination in the context of a two-phase community prevalence study. *Neuroepidemiology*, *16*(3), 141–148.
Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.
Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.
Buschke, H. (1984). Cued recall in amnesia. *Journal of Clinical Neuropsychology*, *6*, 433–440.
Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
Byrne, B.M. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*, 155–175.
Canadian Study of Health and Aging Working Group (1994). Canadian Study of Health and Aging: Study methods and prevalence of dementia. *Canadian Medical Association Journal*, *6*, 433–440.
Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492.
Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.
Cortina, J.M. (2002). Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management*, *28*, 339–362.
Dudgeon, P.L. (2004). A note on extending Steiger's (1998) multiple sample RMSEA adjustment to other noncentrality parameter-based statistics. *Structural Equation Modeling*, *11*, 305–319.
Flanagan, D.P. & Harrison, P.L. (2005). *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.). New York: The Guilford Press.
Gladsjo, J.A., McAdams, L.A., Palmer, B.W., Moore, D.J., Jeste, D.V., & Heaton, R.K. (2004). A six-factor model of cognition in schizophrenia and related psychotic disorders: Relationships with clinical symptoms and functional capacity. *Schizophrenia Bulletin*, *30*(4), 739–754.
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.
Hu, L. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453.
Marsh, H.W., Hua, K-T., & Wen, Z. (2004). In search of golden-rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.
Meredith, W. & Teresi, J.A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(Suppl 3), S69–S77.
Muthen, L.K. & Muthen, B.O. (1998–2007). *Mplus User's Guide, Fifth Edition*. Los Angeles, CA: Muthen & Muthen.
Omura, K. & Sugishita, M. (2004). Simultaneous confirmatory factor analysis of the Wechsler Memory Scale–Revised for two standardization samples: A comparison of groups from Japan and the United States. *Journal of Clinical and Experimental Neuropsychology*, *26*, 645–652.
Preacher, K.J. & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13–43.
Ramirez, M., Teresi, J.A., Holmes, D., Gurland, B., & Lantigua, R. (2006). Differential item functioning (DIF) and the Mini-Mental State Examination (MMSE): Overview, sample, and issues of translation. *Medical Care*, *44*. S95–S106.
Reilly, R.R., Bowden, S., Bardenhagen, F.J., & Cook, M.J. (2006) Invariance of the measurement model underlying depressive symptoms in patients with temporal lobe epilepsy. *Journal of Clinical and Experimental Neuropsychology*, *28*, 1–15.
Rey, A. (1964). *L'examen clinique en psychologie*. Paris, France: Presses Universitaires de France.
Rosen, W.G. (1980). Verbal fluency in aging and dementia. *Journal of Clinical Neuropsychology*, *2*, 135–146.
Satz, P. & Mogel, S. (1962). An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology*, *18*, 77–79.
Spreen, O. & Benton, A.L. (1977). *Neurosensory center comprehensive examination for aphasia*. Victoria, BC: University of Victoria.
SPSS Inc. (2005). *SPSS 14.0 for Windows*. (2005). Chicago, III: SPSS Incorporated.
Strauss, M. & Smith, G.T. (in press). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology* in press.
Teng, E.L. & Chui, H.C. (1987). The Modified Mini-Mental State (3MS) examination. *Journal of Clinical Psychiatry*, *48*, 314–318.
Teresi, J.A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(Suppl 3), S39–S49.
Tuokko, H., Kristjansson, E., & Miller, J.A. (1995). The neuropsychological detection of dementia: An overview of the neuropsychological component of the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology*, *17*, 352–373.
Tuokko, H., Vernon-Wilkinson, R., Weir, J., & Beattie, B.L. (1991). Cued recall and early identification of dementia. *Journal of Clinical and Experimental Neuropsychology*, *13*, 871–879.
Tuokko, H. & Woodward, T. (1996). Development and validation of the demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology*, *18*, 479–616.
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.
Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology*, *19*, 87–95.
Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised Manual*. New York: The Psychological Corporation.
Whitely, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.
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.