Bulletin of Special Education
2015, 40(2), 91-110
DOI: 10.6172/BSE.201507.4002004

Validation of Taiwan WISC-IV by Using Four- and Five-Factor Interpretative Approaches

Hsin-Yi Chen
Professor,
Dept. of Special Education,
National Taiwan
Normal University

Li-Yu Hung
Professor,
Dept. of Special Education,
National Taiwan
Normal University

Yung-Hua Chen
Professor,
The Chinese Behavioral
Science Corporation

Jianjun Zhu
Clinical Assessment,
Pearson

Timothy Z. Keith
Professor,
Dept. of Educational Psychology,
The University of Texas at Austin

Purpose: Invariance is a fundamental property of any instrument used to compare individuals from subpopulations. In empirical settings, the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV) is frequently employed as a part of psychoeducational assessments. Implicit in this common practice is the assumption that WISC-IV scores have the same meaning for children in various subpopulations. The current WISC-IV manual recommends a four-factor scoring structure, whereas the Cattell-Horn-Carroll theory-based five-factor approach recommends that WISC-IV subtests assess five, instead of four, meaningful latent broad factors under *g*. This study investigated the factorial invariance of both four- and five-factor approaches among large normative and mixed exceptional children samples in Taiwan. **Methods:** Data from two large and reliable samples were analyzed. The normative sample was part of the Taiwan WISC-IV standardization sample, which consisted of 704 children aged 9 to 16 years. The overall mean full-scale intelligent quotient (FSIQ) was 100.1 (*SD* = 15.2). The sample consisting of excep-

* Corresponding Author: Hsin-Yi Chen (hsinyi@ntnu.edu.tw)。

tional children was a heterogeneous sample that included 697 children in the special education system with various diagnoses such as intellectual disability, autism, learning disabilities, attention deficit hyperactivity disorder, and other emotional and behavioral disturbances. The mean FSIQ in this sample was 84.9 (*SD* = 19.3). Tests for the higher-order confirmatory factor invariance among the normative and exceptional children samples were based on an analysis of the mean and covariance structure models. Both higher-order WISC-IV four- and five-factor approaches were analyzed. **Results/Findings**: The results based on the multigroup mean and covariance structure analysis revealed the following: (a) Both the four- and five-factor models provided a good data fit for children in both samples, suggesting that both models provide meaningful strategies for interpreting WISC-IV scores. (b) Both models demonstrated full factorial invariance between normative and exceptional samples. (c) Arithmetic, Similarities, and Symbol Search subtests were found with cross-loadings. **Conclusions/Implications:** For both the four- and five-factor approaches, the WISC-IV subtests demonstrate the same underlying theoretical latent constructs, the same strength of relationships among factors and subtests, the same validity of each first-order factor, and the same communalities, regardless of clinical status. The results support the same interpretive approach and meaningful comparisons of the WISC-IV between normative and exceptional children in Taiwan. In addition, when performance inconsistencies for subtests within the same latent ability dimension are detected, or when examiners wish to test specific hypotheses, both the main and minor sources of influence for Arithmetic, Similarities, and Symbol Search subtests may warrant consideration.

Keywords:multigroup mean and covariance structure analysis (MG-MACS), factorial invariance, Wechsler scales

Introduction

Wechsler tests are among the most widely used intelligence instruments worldwide (Archer et al., 2006; Bowden, 2013; Georgas et al., 2003). “The Wechsler Intelligence scales are considered to be among the best of all psychological tests because they have sound psychometric properties and produce information relevant to practitioners” (Groth-Marnat, 2009). Invariance is a fundamental property of any instrument that may be used to compare individuals from subpopulations (Drasgow, 1984, 1987; Millsap & Kwok, 2004; Vandenberg & Lance, 2000). Meaningful comparisons can be made only if the measures are comparable (Chen, Sousa, & West, 2005). In empirical settings, the Wechsler Intelligence Scale for Children , Fourth Edition (WISC-IV; Wechsler, 2003a, 2007) is frequently employed as a part of psychoeducational assessments (Flanagan & Kaufman, 2004; Rabin, Barr, & Burton, 2005; Sattler & Dumont, 2004; Prifitera et al., 2005; Weiss et al., 2008). Implicit in this common practice is the assumption that WISC-IV scores have the same meaning for children in various subpopulations. Thus, demonstrating that WISC-IV scores have the same meaning for children in different subpopulations is crucial.

The contemporary science of intelligence typically agrees on a hierarchical model of cognitive abilities. General intelligence (*g*) tends to emerge whenever a sufficient number of cognitively complex variables are analyzed (Carroll, 1993). In addition to the higher-order *g*, two approaches are currently suggested in the literature for WISC-IV interpretations: four- and five-factor structures. The current WISC-IV manual recommends a four-factor scoring structure: Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI), and Processing Speed (PSI), whereas the Cattell-Horn-Carroll theory[1] (CHC theory, Carroll, 1993, 2005)-based five-factor approach recommends that WISC-IV subtests assess five, instead of four, meaningful latent broad factors under *g*: crystallized intelligence (Gc), visual processing (Gv), fluid reasoning (Gf), short-term memory (Gsm), and processing speed (Gs). Even if the terminology used in CHC and WISC-IV differs in the literature, the underlying constructs of many abilities are similar: For example, both Gc and VCI cover verbal abilities that are accumulated over time and more culturally loaded; Gsm and WMI both cover the short-term and working memories; and Gs and PSI both refer to general speediness. Compared with the four-factor model, the five-factor approach further splits the perceptual reasoning subtests in PRI into two purer subfactors: Gf and Gv.

Studies worldwide have shown firm support for the WISC-IV four-factor scoring structure; it has been confirmed to have good model-data fit for normative samples across cultures (Chen et al., 2009; Georgas et al., 2003; Keith et al., 2006), and for students with clinical diagnoses (Bodin et al., 2009; Watkins et al., 2006). Chen et al (2010) found the four-factor model invariant across four distinct Asian cultures represented by samples from Hong Kong, Macau, Taiwan, and mainland China. Watkins (2006) reported that the four-factor model exhibited a good fit with U.S. data, but recommended that practitioners not discount the strong general factor by overemphasizing the

interpretation of the four first-order factors. Watkins et al. (2006) found the four-factor model also fit well in students with special needs, and suggested the WISC-IV factor structure in clinical sample mirrored that of the standardization sample. Other researchers successfully replicated the four-factor structure in independent, large, and heterogeneous neuropsychological samples which include ADHD, autism and Asperger's, learning disorders, intellectual Disabilities, bipolar disorder, and many more other diagnoses (Bodin et al., 2009; Devena et al., 2013). Individual differences among the four factor-based index scores are considered clinically important, which help to determine relative strengths and weaknesses for children with various clinical diagnoses (Prifitera et al., 2008). For example, children with mathematics disorder seem to show a pattern of PRI<VCI, while the opposite pattern of PRI>VCI was found for children with reading disorder; Children with autistic spectrum disorders were found with relative weakness in WMI and PSI among the four factors, and children with motor impairment demonstrated relative strengths in VCI and weakness in PRI and PSI (Wechsler, 2003b). Many evidences revealed that children with various clinical diagnoses show unique cognitive patterns among these four factors. Advanced investigation on WISC-IV four- and five-factor structure thus is worthy of study.

Keith et al. (2006) analyzed U.S. WISC-IV norming data, and argued that, although the four-factor model fit the data well, the five CHC broad abilities exhibited a superior fit. In addition, several subtests were found to measure multiple abilities, and show possible cross-loadings. Chen, Keith, Chen, and Chang (2009) investigated the construct validity of both the four- and five- factor approaches by using a Taiwan norming sample, results revealed that both models provided meaningful explanations, and the variance explained was similar.

No study has evaluated the clinical validity of both WISC-IV four- and five-factor models until recently. Because WISC-IV is pervasively used in psychoeducational evaluations setting, recent international research efforts were advanced with the exploration of whether WISC-IV factorial invariance holds for normative and clinical samples. Invariance analyses based on a large U.S. data set reported clinical invariance in both solutions. Chen and Zhu (2012) found that the four-factor model exhibited invariance between a mixed clinical and a nonclinical sample of U.S. children. Only the intercept of the Coding and Comprehension subtest varied slightly, and no significant differences in factor patterns or loadings emerged. Following this line of research, Weiss, Keith, Zhu, and Chen (2013) explored the best-fitting U.S. four- and five-factor models by allowing subtests with cross-loadings, and found a nearly full factorial invariance for both models across large normative and clinical samples, with only one subtest error variance discrepancy identified. These above U.S. results have suggested that WISC-IV index scores and subtests have the same meaning for children in both normative and clinical groups.

In Taiwan, the factorial invariance of WISC-IV across large normative and clinical samples has not been investigated. In the literature, the best-fitting four- and five-models identified in Taiwan (Chen et al., 2009) and the U.S. (Keith et al., 2006; Weiss et al., 2013) differed slightly,

different sets of subtests were found with cross-loadings. Thus, factorial invariance research based on the models identified locally is in required to validate a meaningful WISC-IV interpretation for children in Taiwan. This study investigates invariance with large Taiwan samples with considerable variations. Specifically, based on the best-fitting four- and five-factor models identified in Taiwan, we further this line of research by evaluating whether the WISC-IV subtests measure latent abilities in the same manner for the normative sample as they do for children with special needs in Taiwan.

Method

Participants

We analyzed data from two large and reliable samples: The normative sample was based on the Taiwan WISC-IV standardization sample (Wechsler, 2007). The sample with exceptional children was a heterogeneous sample collected from the special education diagnostic system in Taipei City during 2011–2012. All exceptional children were evaluated using multiple evaluation procedures, and the Department of Education, Taipei City Government, formally identified their special needs.

To match the background characteristics and the sizes of both the normative and the exceptional children samples, the normative sample consisted of 704 children aged 9 to 16 years. It was part of the 968 children recruited to standardize the Taiwan WISC-IV. This norm sample was stratified carefully to match the 2006 Taiwan Census for demographics such as gender, parental education, and region. The overall mean Full-scaled IQ (FSIQ) was 100.1 (*SD* = 15.2). For all subtests, the means ranged from 9.96 to 10.05, standard deviations from 2.97 to 3.19, skewness from -0.57 to 0.10, and kurtosis from -0.30 to 0.69. The mean age was 12.9 years (*SD* = 2.3).

The heterogeneous exceptional-children sample included 697 children with various diagnoses in the special education system. These children had a mild or moderate intellectual disability (16.6%); autism or Asperger syndrome (19.8%); learning disabilities (53.4%); attention deficit hyperactivity disorder (ADHD) or ADHD with learning disabilities (8.9%); and other emotional and behavioral disturbances (1.3%). The mean FSIQ in this sample was 84.9 (*SD* = 19.3). For all subtests, the descriptive statistics were as follows: means, 6.63–8.39; standard deviations, 3.36–4.22; skewness, -0.39–0.46; and kurtosis, -0.90–0.38. The age range of 9 to 16 years and the average age (i.e., 12.9 years) were similar to that of the normative sample (*SD* = 1.0).

The relatively equal sample sizes between 2 samples were designed to prevent excessive power in either group, and thus, ensured that the results would not be heavily weighted toward either sample. The demographics of all of the studied samples are listed in Table 1. The mean age and FSIQ are presented, followed by the percentages of sample representation according to gender.

Instrumentation

The Taiwan WISC-IV has 10 core subtests (Similarities [SIM], Vocabulary [VOC], Comprehension [COM], Block Design [BLD], Picture Concepts [PCn], Matrix Reasoning [MR], Digit Span [DS], Letter-Number Sequencing [LNS], Coding [CD], Symbol Search [SYS]) and four

Table 1 Demographic data for the studied samples.

	Normative sample	Clinical Sample (Overall)	Clinical sub-samples				
			ID Mild/Moderate	AUT/ ASP	LD	ADHD	OEBD
N	704	697	116	138	372	62	9
Age							
M	12.9	12.9	12.6	12.4	13.2	13.4	14.1
SD	2.3	1.0	0.9	0.9	0.8	0.8	0.7
FSIQ	100.1	84.9	56.2	90.9	89.4	97.0	94.0
M	15.2	19.3	9.3	24.0	10.5	13.5	13.5
SD							
Gender %							
Female	49.9	24.2	44.8	15.9	23.4	8.1	33.3
Male	50.1	75.8	55.2	84.1	76.6	91.9	66.7

Note: ID Mild = Intellectual Disabilities-Mild Severity; ID Moderate = Intellectual Disabilities-Moderate Severity; AUT = Autistic Disorder; ASP = Asperger's Disorder; LD = Learning Disabilities; ADHD = Attention-Deficit/Hyperactivity Disorder; OEBD = other emotional and behavioral disturbances

supplemental subtests (Information [INF], Picture Completion [PIC], Arithmetic [AR], and Cancellation [CA]). All composites and subtests have demonstrated good reliability, with average internal reliability estimates ranging from 0.85 to 0.96 for composites, 0.74 to 0.91 for core subtests, and .77 to .86 for supplemental subtests (Wechsler, 2007). We employed 10 core subtests and three supplemental subtests in this study to ensure adequate markers for as many latent abilities as possible. We excluded the supplemental Cancellation subtest because the majority of the exceptional children did not take this optional subtest. We considered excluding the Cancellation test to not jeopardize the meaning of the current research findings for the following reasons: (1) Cancellation was confirmed to be a pure processing speed subtest with no cross-loadings (Weiss et al., 2013; Chen & Zhu, 2012); and (2) The latent processing speed factor can be identified successfully by referring to both the CD and SYS subtests in this study.

Analysis

Tests for the higher-order confirmatory factor structure were based on an analysis of the mean and covariance structure models, for which we used LISREL version 8.8 (Jöreskog & Sörbom, 2006). We first checked the normality of each subtest. In both groups, skewness ranged from -.57 to .46, and kurtosis ranged from -.90 to .69. Maximum likelihood estimation is known for robustness (Hu & Bentler, 1998), and is considered adequate for data with a skewness of less than 2 and a kurtosis of less than 7 (West et al., 1995). Thus, we used maximum likelihood estimation for model estimation. Interested readers may contact the corresponding author for the correlation matrices.

We tested both the four- and five-factor models with hypothesized cross-loadings verified by Chen et al. (2009). The previously researched best-fitting four-factor model specified a higher-order *g* and four first-order factors. There are four

verbal comprehension subtests (SIM, VOC, COM, and INF) on the first factor, four perceptual reasoning subtests (BLD, PCn, MR, and PIC) on the second factor, three working memory subtests (DS, LNS, and AR) on the third factor, and two processing speed subtests (CD, SYS) on the fourth factor. We allowed the arithmetic subtest to be cross-loaded on the WMI and VCI. The four-factor structure is depicted in Figure 1.

The previously researched best-fitting five-factor model specified a higher-order *g* and five first-order factors. This model specified four subtests (SIM, VOC, COM, and INF) on the verbal comprehension factor (Gc/VCI), two subtests (BLD and PIC) on the visual processing factor (Gv), two subtests (PCn and MR) on the fluid reasoning factor (Gf), three subtests (DS, LNS, and ARI) on the working memory factor (Gsm/WMI), and two subtests (CD and SYS) on the processing speed factor (Gs/PSI). We cross-loaded three subtests: SIM on Gc and Gf, ARI on Gsm and Gc, and SYS on Gs and Gv. The five-factor structure is displayed in Figure 2.

We examined the factorial invariance of both models, and tested six levels of nested models to investigate the degree of invariance (Keith, 2014; Vandenberg, 2002; Wicherts & Dolan, 2010). Each level had more constraints than those of the previous level (Keith & Reynolds, 2012; Meredith, 1993). The initial and weakest level was configural invariance, which assumed the same number of factors and the same overall factor pattern across groups. The second level was first-order factor-loading invariance (or metric/weak factorial invariance). Loadings of subtests on factors were constrained so that factor loadings were equal across groups. When the factor loadings are equal, the scales of the latent variables are the same for both groups, and the unit of measurement is identical. The third level was intercept invariance (or scalar/strong factorial invariance). At this level, any group difference in subtest means result from the true mean differences in latent factors. The subtests have the same intercepts across groups if they have the same latent factor means. The fourth level tested residual invariance (or strict factorial invariance) to examine whether "all group differences on the measured variables are captured by, and attributable to, group differences on the common factors" (Widaman & Reise, 1997). These residuals are a combination of subtest-specific unique variance and measurement errors. The fifth level was second-order factor-loading invariance. We assumed that first-order latent factors show the same amount of change in each group for the same increase in *g*. Finally, we tested the invariance of disturbances (factor unique variances) of the first-order factors. Although disturbance invariance is not fundamentally crucial for measurement invariance, it provides substantial information regarding human cognitive abilities across groups. We did not constrain first-order factor intercepts to be equal across groups, because such constraints addressed measurement questions that do not pertain to the current study. For all analyses, we identified the scale of latent factors by fixing a factor loading of each factor to one. The equality constraints imposed across groups in each level are listed in Table 2.

Multiple indices of the model fit were used to evaluate and compare the models (Bentler & Bonett, 1980; Hoyle & Panter, 1995; Hu & Bentler, 1998, 1999; Kline, 2010; Marsh, Balla, &

Table 2 Equality constraints imposed across groups in each level

Level	Description	Factor loadings		Intercepts	Residual variances	Disturb-ances	1st-order Factor means
		1st-order	2nd-order				
1	Configural invariance	free	free	free	free	free	fixed at 0 for both groups
2	Metric invariance (1st-order)	invariant	free	free	free	free	fixed at 0 for both groups
3	Intercept invariance	invariant	free	invariant	free	free	fixed at 0 for one group only
4	Residual invariance	invariant	free	invariant	invariant	free	fixed at 0 for one group only
5	2nd -order loadings equal	invariant	invariant	invariant	invariant	free	fixed at 0 for one group only
6	1st-order unique variances equal	invariant	invariant	invariant	invariant	invariant	fixed at 0 for one group only

Note. Each step is nested under the previous one; free: freely estimated within each group; invariant: parameters estimated equally across groups

McDonald, 1988; McDonald & Ho, 2002). Single models were jointly evaluated by using the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). An RMSEA value less than .05 corresponded to a good fit, and .08 was considered to be acceptable. SRMR values less than .08 were considered to be good. A value of .95 served as the cutoff point for an acceptable fit of all indices ranging from 0 to 1, with 1 indicating a perfect fit. Change in the chi-square ($\Delta\chi^2$) value was used to evaluate competing nested models (Bentler & Bonett, 1980). The Akaike information criterion (AIC) and sample size adjusted Bayesian Information Criterion (aBIC) were used for comparisons of competing nested and non-nested models (Kaplan, 2000; Loehlin, 2004), with lower values indicating a superior fit. The aBIC has a more substantial reward for parsimony compared with the AIC.

To determine evidence of invariance, consensus is scant regarding the most appropriate criterion (Byrne & Stewart, 2006). Two perspectives were jointly evaluated: (a) the traditional perspective based on $\Delta\chi^2$, and (b) the practical perspective based on differences in the comparative fix index CFI (ΔCFI). Comparatively, the $\Delta\chi^2$ test is known to be oversensitive to the sample size and discrepancies from normality (Kline, 2010; West, Finch, & Curran, 1995). Cheung and Rensvold (2002) recommended ΔCFI as superior to $\Delta\chi^2$ for its independence in model complexity, sample size, and overall fit measures. "A value of ΔCFI smaller than or equal to -.01 indicates that the null hypothesis of invariance should not be rejected" (Cheung & Rensvold, 2002). An absolute ΔCFI value higher than .01 (i.e., |ΔCFI| > .01) was proposed as an indicator of a meaningful fall in fit. Given the large sample sizes, large modeled variables, heterogeneous samples, and the number of comparisons being made in this study, we decided to evaluate the invariance by $\Delta\chi^2$ and ΔCFI

jointly to secure meaningfulness and prevent any unnecessary oversensitivity. The criterion for rejecting the null hypothesis of invariance was set as a P value of less than .001 for the $\Delta\chi^2$ test and an absolute ΔCFI value higher than .01.

Results

Invariance between normative and exceptional-children samples in the four-factor model

Table 3 lists the invariance analyses for the four-factor model. The baseline model fit was first checked for each sample[2]. The model fit each datum well, suggesting that the following invariance verification was meaningful. Variance-covariance matrices were constrained to be equal across groups[3] (Model 1). This constrained model fit the data well (CFI = .99; RMSEA = .055), suggesting fairly invariant WISC-IV subtest covariance patterns in children. Because any factor structure is derived from these variance-covariance matrices, if the WISC-IV measures the same constructs, factor structure between the normative and exceptional samples should be similar.

First, the configural model (Model 2) provided an acceptable fit to the data. Normative and exceptional children shared the same WISC-IV first- and second-order four-factor patterns and the corresponding subtests loaded on the same factors. With the factor pattern established, we imposed cross-group constraints on the first-order factor loadings. The fit did not decrease substantially (Model 3). Although $\Delta\chi^2$ suggested a worse fit than did the configural model, the ΔCFI value was 0, implying that the subtests generally involve measuring the same latent factors in both groups. Next, we constrained the subtest intercepts to be equal (Model 4). To identify this model properly, we fixed the means of the first-order factors in the normative group to zero, but freed those in the exceptional group. Thus, the factor means for the exceptional group represent the mean differences. All corresponding first-order intercepts were constrained to be equal. There was no deterioration of fit with these constraints by ΔCFI. Next, when the subtest residuals were constrained to be equal across groups (Model 5), the addition of subtest residual variance constraints reduced the fit significantly according to $\Delta\chi^2$, but again, not according to ΔCFI. Next, when structural parameters (second-order loadings and first-order unique variances) were constrained as to be equal between groups in steps (Models 6 and 7). There was no result in practical deterioration of fit by ΔCFI.

Because of the complexity of the model and the strictness of the test, we concluded that the WISC-IV exhibits acceptable levels of invariance among four factors between the normative and exceptional-children groups. Differences in subtest scores on the WISC-IV are generally caused by latent constructs, and the test is not biased based on the clinical status. We fixed the means of the four latent factors in the normative group to zero, and the non-standardized latent means for VCI, PRI, WMI, and PSI in the exceptional group were estimated freely as -2.18, -1.88, -2.31, and -3.12, respectively. This finding suggests that, the mixed clinical group scored lower on the underlying four first-order factors compared with the normative sample. The largest discrepancy emerged for the processing speed factor, with a

mean difference over one standard deviation.

Standardized estimates based on Model 7 for both groups are shown in Figure 1. All 13 subtests loaded strongly on the corresponding factors. Consistent with the literature, Arithmetic was confirmed as a mixed measure of the WMI and VCI (factor loadings were .56 and .31, respectively). Across all four first-order factors, PRI had the highest *g* loading (.92). All parameter estimates were reasonable and theoretically sound. Most important, these estimates were found invariant across groups

Table 3 Invariance analyses of four-factor model of normative and exceptional children sample.

Model	χ^2	df	CFI	RMSEA	RMSEA 90%CI	SRMR	AIC	aBIC	Model Comparison	ΔCFI	$\Delta\chi^2$	Δdf	p
Phase I : Baseline model fit for each group													
Normative children (n = 704)	134.1	60	.99	.042	.032,.051	.027	196.10	238.92					
Exceptional children (n = 697)	234.4	60	.99	.065	.056,.073	.029	296.40	338.92					
Phase II : Measurement Invariance across groups													
Model 1 Equality of variance-covariance matrices	285.89	91	.99	.055	.048, .063	.220	467.89	656.11					
Model 2 configural	368.50	120	.99	.054	.048,.061	.029	544.50	726.51	-----	----	----	----	----
Model 3 first-order loadings	420.43	130	.99	.057	.051,.063	.043	576.43	737.76	3 vs. 2	.00	51.93	10	.000
Model 4 first-order loadings and subtest intercepts	586.76	139	.99	.068	.062,.074	.056	724.76	867.47	4 vs. 3	.00	166.33	9	.000
Model 5 first-order loadings, subtest intercepts, and subtest residual variances	617.38	152	.98	.066	.061,.072	.055	729.38	845.21	5 vs. 4	.01	30.62	13	.004
Model 6 first-order loadings, subtest intercepts, residual variances, and second-order loadings	726.67	156	.98	.072	.067,.078	.210	830.67	938.22	6 vs. 5	.00	109.29	4	.000
Model 7 first-order loadings, subtest intercepts, residual variances, second-order loadings, and disturbances of first-order factors	788.12	160	.98	.075	.070,.080	.220	884.12	983.40	7 vs. 6	.00	61.45	4	.000

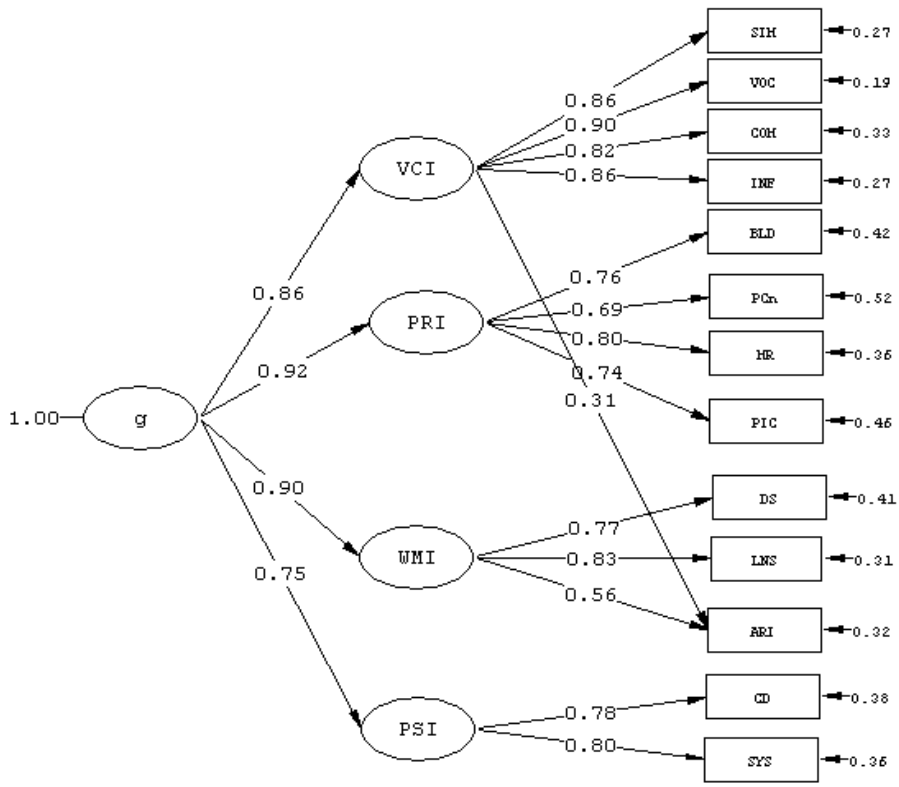

Figure 1 Invariant standardized estimation for the final validated four-factor model for both groups (Model 7 in table 3).

Invariance between normative and exceptional-children samples in the five-factor model

Following a similar procedure, we assessed the factorial invariance of the five-factor model[4]. The fits of the various models are listed in Table 4. When we considered $\Delta\chi^2$ and ΔCFI jointly, the five-factor model demonstrated good levels of factorial invariance between children in both groups. The results supported full measurement invariance.

When the means of the five latent factors in the normative group were fixed to zero, the non-standardized latent means for Gc, Gv, Gf, Gsm, and Gs in the special education group were freely estimated as -1.71, -1.83, -1.62, -2.30, and -3.40, respectively. Again, the finding suggested that, on average, the mixed exceptional group perform lower for the underlying five factors compared with the normative sample; their processing speed tended to decline the most.

Standardized estimates based on Model 7 for both groups are shown in Figure 2. All 13 subtests loaded strongly on the corresponding factors. Consistent with the literature, Arithmetic was confirmed a mixed measure of Gsm (WMI) and Gc (VCI), factor loadings were .54 and .34, respectively; Similarities loaded mainly on Gc (VCI) and minor on Gf (factor loadings were .66 and .22,

respectively); and Symbol Search measures mainly Gs (PSI), with some additional Gv (factor loadings were .57 and .26, respectively). As expected, Gf had the highest *g* loading, suggesting that fluid reasoning is statistically indistinguishable from *g*. All parameter estimates were reasonable and theoretically sound. These estimates suit interpretations for both normative and clinical children.

Table 4 Invariance analyses of five-factor model of normative and exceptional children sample.

Model	χ^2	df	CFI	RMSEA	RMSEA 90%CI	SRMR	AIC	aBIC	Model Comparison	ΔCFI	$\Delta\chi^2$	Δdf	p
Phase I : Baseline model fit for each group													
Normative children (n = 704)	113.90	57	.99	.038	.028,.048	.026	181.90	228.87					
Exceptional children[3] (n = 697)	248.67	57	.99	.070	.061,.078	.031	310.67	353.19					
Phase II : Measurement Invariance across groups													
Model 1 Equality of variance-covariance matrices	285.89	91	.99	.055	.048, .063	.220	467.89	656.11					
Model 2 configural	362.56	114	.99	.056	.049,.062	.031	550.56	744.98	-----	----	----	----	----
Model 3 first-order loadings	413.96	125	.99	.058	.051-.064	.042	580.56	751.63	3 vs. 2	.00	51.40	11	.000
Model 4 first-order loadings and subtest intercepts	556.52	133	.99	.067	.062,.073	.051	706.52	861.64	4 vs. 3	.00	142.56	8	.000
Model 5 first-order loadings, subtest intercepts, and subtest residual variances	594.15	146	.99	.066	.061,.072	.050	718.15	846.39	5 vs. 4	.00	37.63	13	.000
Model 6 first-order loadings, subtest intercepts, residual variances, and second-order loadings	727.81	151	.98	.074	.069,.079	.210	841.81	959.70	6 vs. 5	.01	133.66	5	.000
Model 7 first-order loadings, subtest intercepts, residual variances, second-order loadings, and disturbances of first-order factors	771.66	156	.98	.075	.070,.080	.220	875.66	983.21	7 vs. 6	.00	43.85	5	.000

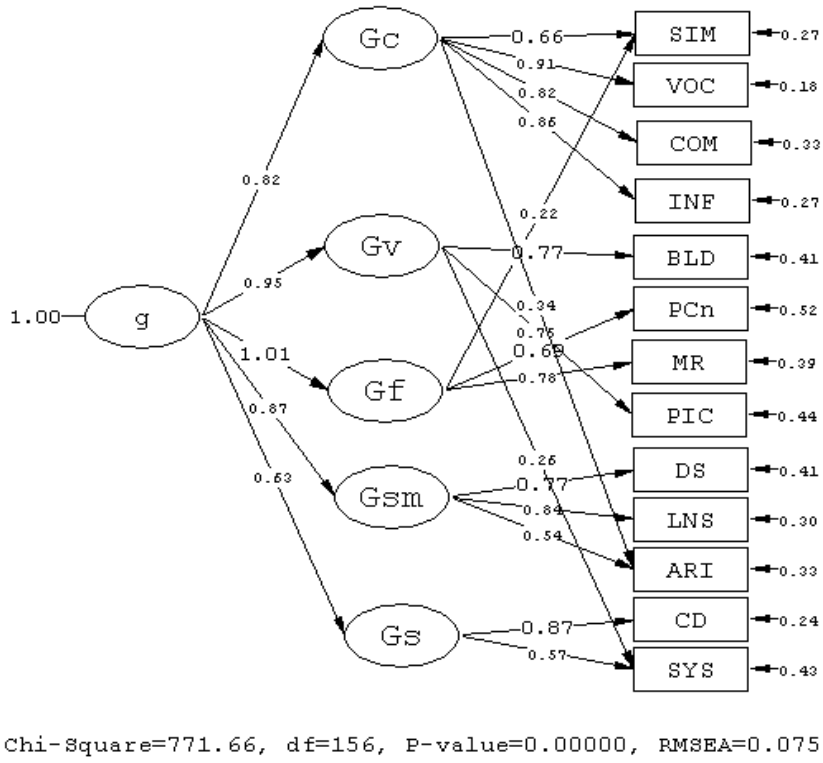

Figure 2 Invariant standardized estimation for the final validated five-factor model for both groups (Model 7 in table 4)

Discussion

We conducted this study to determine the invariance of WISC-IV constructs across large normative and exceptional samples in Taiwan. This study is crucial because it is the first to evaluate the WISC-IV measurement validity of an alternative five-factor interpretive approach in a clinical-children sample from Taiwan.

The first major set of findings is that both the four- and five-factor models fit the data well. Both models provided meaningful strategies for interpreting WISC-IV scores for both normal and exceptional children in Taiwan. Our results supported the notion that both models have psychometric merit and are sound for interpretation.

The second and most critical set of findings is that both the four- and five-factor models derived from the normative data provided a good fit with the exceptional sample data. Each model demonstrated full factorial invariance between normative and exceptional samples. The WISC-IV subtests demonstrate the same underlying theoretical latent constructs, the same strength of relationships among factors and subtests, the same validity of each first-order factor, and the same communalities, regardless of clinical status, thus supporting meaningful comparisons of WISC-IV between these 2 groups in Taiwan. Irrespective of

the model used, the WISC-IV subtests have the same meaning, regardless of the group status.

The results also revealed that the five-factor approach exhibited a slightly superior fit for the normative sample according to both the AIC and the aBIC. By contrast, the four-factor structure seemed to exhibit a slightly superior fit for the mixed exceptional-children sample. Though the differences were indeed trivial, this part of result differs from a previous U.S. finding, which showed that the five-factor approach exhibited a relatively superior fit for both the normative and clinical samples (Weiss et al., 2013). The main difference between the four- and five-factor models is that the perceptual reasoning subtests in the four-factor model are separated into visual processing (Gv) and fluid reasoning (Gf) in the five-factor model. In both the U.S. and Taiwan studies, children with various clinical diagnoses were grouped into large mixed clinical sample, but the number of clinical subtypes involved and the percentage of each category varies between studies. The advantage in such a large sample was that it represents the special student population more closely. However, the disadvantage of the heterogeneous nature of this mixed sample, which may have weakened the meaningfulness in separating Gv and Gf from PRI, could not be avoided. We suspected that children with some specific diagnostic types may exhibit the pattern Gv > Gf, whereas the other subtypes may exhibit a reversed Gv < Gf. All of these subgroup discrepancies may have been confounded when mixed as one overall exceptional sample, thus revealing a slightly superior fit for the four-factor model. Based on the literature, children with an intellectual disability seem to exhibit a slight Gf > Gv pattern, whereas children with ADHD, disruptive behavior, language disorder, and autistic disorder or Asperger syndrome seem to exhibit the opposite (i.e., a Gv > Gf pattern; Wechsler, 2012). To clarify the specific clinical utility and practical applications, we encourage future investigations on further invariance and validation based on specific diagnostic grouping.

A third set of major findings concerns the verification of multiple abilities, as required by some subtests as invariance across both samples. Chen et al (2009) reported the mixed loadings of the Arithmetic, Similarities, and Symbol Search subtests for normal children in Taiwan. Current findings further demonstrated that these previously identified cross-loadings exist, not only for the normative sample but also for the clinical-children sample. When interpreting the WISC-IV result for children in Taiwan, the performance in the Arithmetic subtest should be considered to be influenced chiefly by working memory and some verbal comprehension; the performance in the Similarities subtest should be considered to be influenced chiefly by verbal comprehension and some fluid reasoning; and the performance in the Symbol Search subtest should be considered to be influenced chiefly by the processing speed and some visual processing. Such an understanding is valuable for practitioners in the field of assessment. When performance inconsistencies for subtests within the same latent ability dimension are detected, or when examiners wish to test specific hypotheses, both the main and minor sources of influence for these three subtests may warrant consideration.

Our findings have crucial applied implications. The acceptability of the four- and five-

factor models in both samples suggested that they are complementary models for interpreting WISC-IV findings. For children with consistent subtest scores within each of the four WISC-IV composites, the current four factors provide a parsimonious solution, and constitute an appropriate level of interpretation. For children with discrepant subtest scores within some of the four composites, the five-factor model suggests a likely interpretive reorganization. For children who present a subtest scattered within the PRI composite, the findings suggested that a common pattern may be consistent between Gv subtests (i.e., Block Design and Picture Completion) and between Gf subtests (i.e., Matrix Reasoning and Picture Concepts). Furthermore, our data suggested that this interpretative approach should be equally applicable for children from both the general population and that of children with special needs. (Please refer to sophisticated WISC-IV texts such as Sattler and Dumont (2004), Flanagan and Kaufman (2004), and Prifitera et al.(2008) for a detailed discussion on WISC-IV score interpretation suggestions.)

Overall, this study provided practical WISC-IV validity evidence. However, the limitations warrant attention. First, we investigated the factorial invariance by using a mixed clinical sample. Studies based on each clinical subsample may provide further useful group-specific information. Second, the normative and clinical samples were not a perfect match for age variations and gender ratios. We considered this to be harmless for our conclusion based on the known age and gender invariance in the literature (Chen & Zhu, 2008, 2009; Keith et al., 2010). Nonetheless, we recommend that future research be conducted with ideal samples. Finally, we used $\Delta\chi^2$ and ΔCFI jointly as a criterion for rejecting the null hypothesis of invariance. Kenny (2014) indicated that CFI results in a penalty of 1 point for every parameter estimated, and this penalty of only 1 for the model complexity may be too low. We encourage future studies to determine whether this property renders ΔCFI a less sensitive index, especially for models with large samples and numerous estimated parameters. Meade et al (2006) recommended the best practice in test of measurement invariance would be to report the change in all four fit indices: Chi square difference test, McDonald's NCI, CFI, and Gamma-hat. They also suggested that researchers may then use their own judgment if some but not all of the fit indices show a lack of invariance depending upon their own research needs. Future explorations on utilities of various fit indices are encouraged.

Validity evidence from both psychometric and empirical perspectives should be accumulated continuously (AERA, APA, NCME, 2014). We recommend that validation studies based on clinical performance, such as diagnostic differentiation, be conducted in the future. We also encourage future studies that investigate whether five composites present a superior separate special subtype of clinical groups versus the four composites.

Notes

[1] The Cattell-Horn-Carroll model categorizes cognitive abilities into three structural levels. At the top of the hierarchy is general intelligence (*g*). In the middle are 10 broad abilities (crystallized intelligence [Gc], fluid intelligence [Gf], quantitative knowledge [Gq], short-term memory [Gsm], long-term retrieval [Glr], visual processing [Gv],

auditory processing [Ga], processing speed [Gs], reading and writing ability [Grw], and decision-reaction time-speed [Gt]). More than 70 narrow abilities are on the base. This model is still being extended.

[2] For verification, we divided the clinical group into subgroups, and found the degree of the four-factor baseline model fit was adequate/acceptable for those with an intellectual disability (CFI = .96, RMSEA = .063, SRMR = .074); and in the autism spectrum (CFI = .98, RMSEA = .10, SRMR= .044); and LD (CFI = .96, RMSEA = .058, SRMR = .055). The samples for ADHD and OEBD were too small for conducting a separate analysis.

[3] When 「Equality of variance-covariance matrices」are supported across groups, the instrument generally measures the same constructs across groups (without demonstrating what those constructs are). If not supported, the detailed steps for testing invariance are needed to find the source of misfit.

[4] Similarly, we examined the degree of the five-factor baseline model fit in 3 clinical subgroups. The fit was mostly acceptable: Intellectually disabled (CFI = .95, RMSEA = .068, SRMR = .083); autism spectrum (CFI = .97, RMSEA = .12, SRMR= .0505); and LD (CFI = .96, RMSEA = .059, SRMR = .055).

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Archer, R. P., Buffington-Vollum, Stredny, J. K., & Handel, R. W. (2006). A survey of test use patterns among forensic psychologists. *Journal of Personality Assessment, 87,* 64-94. doi: 10.1207/s15327752jpa8701_07

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606. doi: 10.1037//0033-2909.88.3.588

Bodin, D., Pardini, D.A., Burns, T.G., Stevens, A.B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Child Neuropsychology, 15*(5), 417-424. doi: 10.1080/09297040802603661

Bowden, S. C. (2013). Theoretical Convergence in Assessment of Cognition. *Journal of Psychoeducational Assessment, 31*(2), 148-156. doi: 10.1177/0734282913478035

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: a walk through the process. *Structural Equation Modeling A Multidisciplinary Journal*, *13*(2), 287- 321. doi: 10.1207/s15328007sem1302_7

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

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 69-76). New York, NY: Guilford.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second–order factor models. *Structural Equation Modeling A Multidisciplinary Journal, 12*(3),

471-492. doi: 10.1207/s15328007sem1203_7

Chen, H., Keith, T., Chen, Y., & Chang, B. (2009). What does the WISC-IV measure? : Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences,54*(3), 85-108.

Chen, H., Keith, T., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan. *Personality and Individual Differences, 49*(7), 677-682. doi: 10.1016/j.paid.2010.06.004

Chen, H., & Zhu, J. (2008). Factor invariance between genders of the Wechsler Intelligence Scale for Children-Fourth Edition. *Personality and Individual Differences, 45*(3), 260-266. doi: 10.1016/j.paid.2008.04.008

Chen, H., & Zhu, J. (2009). Testing for WISC-III factorial invariance across gender。*Psychological Testing, 56*(1), 1-18.

Chen, H., & Zhu, J. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences, 52*(2), 161-166. doi: 10.1016/j.paid.2011.10.006

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling A Multidisciplinary Journal, 9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Devena, S. E., Gay, C. E., &Watkins, M. W. (2013). Confirmatory factor analysis of the WISAC-IV in a hospital referral sample. *Journal of Psychoeducational Assessment, 31*(6), 591-599. doi: 10.1177/0734282913483981

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95,* 134–135. doi: 10.1037//0033-2909.95.1.134

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29. doi: 10.1037/0021-9010.72.1.19

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment*. Hoboken, NJ: Wiley.

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. San Diego, CA: Academic Press.

Groth-Marnat, G (2009). *Handbook of Psychological Assessment.* (5th ed.). Hoboken, NJ: Wiley.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp.158-176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. doi: 10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 statistical program*. Lincolnwood, IL: Scientific Software.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks,

CA: Sage.

Kenny. D. A. (2014). Measuring model fit. Retrieved August 2, 2014, from http://davidakenny. net/cm/fit.htm

Keith, T. Z. (2014) *Multiple regression and beyond: An introduction to multiple regression and structure equation modeling*(2nd ed). NY: Routledge.

Keith, T.Z., Fine, J.G., Taub, G., Reynolds, M.R., Kranzler, J.H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children – Fourth Edition: What does it measure? *School Psychology Review,* 35 (1), 108-127.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales–II: Consistency across ages 4 to 17. *Psychology in the Schools, 47*(7), 676–697. doi: 10.1002/pits.20498

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 758-799). New York, NY: Guilford.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93–115. doi: 10.1037/1082-989X.9.1.93

Loehlin, J. C. (2004). *Latent variable models: an introduction to factor, path, and structural equation analysis* (4th rev.ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. *Psychological Bulletin, 103*(3), 391-410. doi: 10.1037/0033-2909.103.3.391

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82. doi: 10.1037/1082-989X.7.1.64

Meade, A. W., Johnson, E. C., Braddy, P. W. (2006). The utility of alternative fit indices in tests of measurement invariance. *Academy of Management Proceedings , 2006*(1), B1 - B6. doi: 10.5465/AMBPP.2006.27182124

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. doi: 10.1007/BF02294825

Prifitera, A., Saklofske, D. H., and Weiss, L. G. (Eds) (2008). *WISC-IV clinical assessment and intervention 2e*. San Diego, CA: Academic Press.

Prifitera, A., Saklofske, D. H., & Weiss, L. G.. (Eds) (2005). *WISC-IV clinical use and interpretation: Scientist-Practitioner perspectives*. San Diego, CA: Academic Press.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA division 40 members. *Archives of Clinical Neuropsychology, 20,* 33-65. doi: 10.1016/j.acn.2004.02.005

Sattler, J. M., & Dumont, R. (2004). *Assessment of children: WISC–IV and WPPSI–III supplement.* Le Mesa, CA: Author.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69. doi: 10.1177/109442810031002

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment, 18*(1), 123-125. doi: 10.1037/1040-3590.18.1.123

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor Structure of the Wechsler Intelligence Scale for Children-Fourth Edition Among Referred Students. *Educational and Psychological Measurement, 66*(6), 975–983. doi: 10.1177/0013164406288168

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

Wechsler, D. (2003b). *Wechsler Intelligence Scale for Children-Fourth Edition technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2007). *Wechsler Intelligence Scale for Children-Fourth Edition administration manual (Taiwan version)*. Taipei city, Taiwan, ROC: Chinese Behavioral Science Corporation.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition technical and interpretive manual.* San Antonio, TX: Pearson.

Weiss, L. G., Keith, T., Zhu, J., & Chen, H. (2013): WISC-IV and Clinical Validation of the Four- and Five-Factor Interpretative Approaches. *Journal of Psychoeducational Assessment, 31(2),* 114-131. doi: 10.1177/0734282913478032

Weiss, L G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (Eds) (2008). WISC-*IV advanced clinical interpretation.* San Diego, CA: Academic Press.

West, S. G., Finch, J. P., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage.

Wicherts, J. M. & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and practice, 29*(3), 39-47. doi: 10.1111/j.1745-3992.2010.00182.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281-324). Washington: American Psychological Association.

收稿日期：2014.04.09
接受日期：2015.01.22

特殊教育研究學刊
民 104，40 卷 2 期，91-110 頁

WISC-IV 兩種因素模式解釋取向之效度研究

陳心怡
臺灣師大特教系教授

洪儷瑜
臺灣師大特教系教授

陳榮華
中國行為科學社退休教授

朱建軍
皮爾森測驗公司心理計量部主任

Timothy Z. Keith
美國德州大學奧斯汀校區教心系教授

本研究目的在分析魏氏兒童智力量表第四版（WISC-IV）在一般兒童與特殊兒童組別間之因素恆等性。根據 9-16 歲之台灣標準化樣本（N=704）及身心障礙特殊兒童樣本（N=694），研究者以多樣本高階結構方程模式對「WISC-IV 現行計分之四因素模式」及「Cattell-Horn-Carroll（CHC）理論依據之五因素模式」兩種不同解釋取向進行系列性因素恆等性檢驗。兩種架構的主要差異是，四因素內的「知覺推理」因素在五因素架構內被進一步區分為「流體推理」和「視覺空間」兩個因素。本研究發現：（1）對一般兒童與特殊兒童而言，四因素與五因素模式均為合理且具意義之詮釋方式。算術、類同、與符號尋找分測驗具有跨因素負荷量，表示它們測量到多元認知內涵，分析詮釋時宜納入考量；（2）四因素及五因素模式在兩組兒童間均具因素恆等性，WISC-IV 分數對兩組兒童具相同的構念意義、因素與分測驗關連性、及因素與分測驗效度。研究結果支持一般兒童與特殊兒童的 WISC-IV 分數結果可用相同方式詮釋，不論是四或五因素架構均可被用來對兩組兒童進行有意義之臨床比較。

關鍵字：多樣本高階結構方程模式、測量恒等性、魏氏量表