特殊教育研究學刊
民 97，33 卷 1 期，93-107 頁

WISC-III Factorial Invariance Across Normal and Clinical Samples

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

Zhu Jianjun
Director, The Psychological Corporation, U.S.A.

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

Chang Cheng-Fen
Professor, Dept. of Special Education, National Taiwan Normal University

Yang Tsung-Ren
Associate Professor, Dept. of Special Education, National Taipei University of Education

ABSTRACT

Taking as its sample a normal group of 1,100 children and a clinical group of 1,150 children in Taiwan, the purpose of this study was to test for invariance in the WISC-III factorial structure across normal and clinical samples. Results of our two-stage multi-sample confirmatory factor analysis supported partial factor invariance. Among all examined parameters, only a few discrepancies involving unique error variances and subtly correlated residuals were found across groups. Evidence of partial measurement invariance indicated that the hypothesized four-factor model generally fit the real data fairly well. It was found that while measurement accuracy for a few subtests differs between and among groups, the overall main structure and factor loadings were generally invariant across normal and clinical samples. Therefore we concluded that WISC-III scores can be interpreted as having the same meaning across groups.

Keywords: Exceptional children; Factorial Invariance; Multi-sample CFA; WISC-III

---

*Our gratitude goes to the Academic Paper Editing Clinic, NTNU.

Introduction

The Wechsler Intelligence Scale for Children (WISC) is one of the most widely referred individualized intelligence tests (Camara, Nathan, & Puente, 2000). Up to now, it is estimated that, including Taiwan, close to twenty countries have adapted the standardized version of this instrument (Georgas, Weiss, Van de Vijver, & Saklofske, 2003).

Compared to the traditional verbal － performance Wechsler IQ construct, a new, four-factor structure (Verbal Comprehension, Perceptual Organization, Freedom from Distractibility, and Processing Speed) was proposed in the third edition of this instrument in 1991 (WISC-III; Wechsler, 1991). This four － factor-based structure is more in line with contemporary research on intellectual components and has been recognized for its' clinical utility both in Taiwan and internationally (Chen & Yang, 2000; Hung, Chen, & Chen, 2003; Weiss, Saklofske, Schwartz, Prifitera, & Courville, 2006). Besides, this model has been cross-validated extensively in a variety of samples by traditional exploratory or confirmatory factor analyses (Donders & Warschausky, 1996; Keith & Witta, 1997; Konold, Kush, & Canivez, 1997; Roid, Prifitera, & Weiss, 1993; Roid & Worrall, 1997; Tupa, Wright, & Fristad, 1997) , and was confirmed to be a preferred model for Taiwan normal children (Chen, Zhu, & Chen, 2000; Georgas, Van de Vijver, Weiss, & Saklofske, 2003). Nonetheless, examination of the factorial invariance between normal and clinical populations based on multi-group structure equating modeling (SEM) is still short.

Factorial invariance is a key property of any measure (Drasgow, 1984, 1987; Rock, Werts, & Flaugher, 1978). Scores for individuals from different groups cannot be given the same meaning if there is no evidence of such invariance (Horn & McArdle, 1992; Vandenberg & Lance, 2000). Furthermore, it is clearly stated in the standard 7.8 of "Standards for Educational and Psychological Testing" (AERA, APA, & NCME, 1999) that, "Comparisons across groups are only meaningful if scores have comparable meaning across groups. The standard is intended as applicable to settings where scores are implicitly or explicitly presented as comparable in score meaning across groups (p.83)". Millsap and Kwok (2004) also pointed out that selection based on a composite with underlying non-invariant factor structures could be problematic.

In empirical practice, WISC-III is most frequently administered for the purpose of diagnosing and evaluating the cognitive function and exceptionality of clinical populations (Kaufman, 1994; Sattler, 2001; Sattler & Dumont, 2004; Prifitera, Saklofske, & Weiss, 2005). Implicit in this common practice is the assumption that WISC-III subtests and factors have the same meaning for both normal and clinical children; that is, equivalence is assumed to hold for the underlying theoretical structures, factor patterns (subtests loaded on the same factors across groups) and the magnitudes of factor loadings.

In the literature, few studies ever examine the measurement equivalence of WISC-III across large normal and clinical samples by SEM technique which has the advantage of taking covariance matrix of both groups into consideration. Therefore, taking as its sample Taiwanese children with a sufficiently large sample size and

degree of variation, this study investigated the issue of invariance in the WISC-III by means of a multi-sample structure equating modeling.

Method

Participants

Data based on two samples were analyzed in this study. The first one is a normal sample, and the second one is a clinical sample.

The normal sample

The first sample is the WISC-III Taiwan standardization sample, comprised of the responses of 1,100 normal children ranging in age from 6 to 16 years old. This nationally representative sample was divided into 11 groups according to age, with 50 males and 50 females in each group. This sample was selected carefully to match the Taiwan census information on several variables, including region, gender, and parents' educational level. The mean age was 11, with a standard deviation of 3.16; the average full-scaled IQ (FSIQ) was 100 (SD=15). A detailed description of this normal sample is reported in the Taiwan version of the WISC-III manual (Wechsler, 1997).

The clinical sample

The second sample is a heterogeneous clinical sample including a total of 1,150 Taiwan children, who were formally identified and diagnosed by clinicians or educational evaluators as with special needs. Among them, 37% were diagnosed as being mentally retarded, 32% as having learning disabilities, 19% as being autistic, 10% as having ADHD, and 2% as having emotional and behavioral disorders. The data were collected by authors via multiple tracking methods during 2002 to 2007. Some were from formal academic or clinical evaluation records, and others were from database in special education identification and placement system. For this group of children, the mean FSIQ was 78.59 (SD=21.78). The average age was similar to that of the normal group (M=10.72, SD=2.69), and the gender ratio was roughly 7:3 (72% males vs. 28% females), which concurs with the known fact that there is a much higher percentage of males in the clinical population.

Instrumentation

The Taiwan version of the WISC-III (Wechsler, 1997) contains 13 subtests: Information (INF), Similarity (SIM), Vocabulary (VOC), Comprehension (COM), Picture Completion (PIC), Picture Arrangement (PA), Block Design (BLD), Object Assembly (OA), Arithmetic (ARI), Digit Span (DS), Coding (CD), Symbol Search (SYS), and Mazes (MZ). All composites and subtests demonstrated good reliabilities (ex., the internal reliability ranging from .87 to .96 for composites, and .68 to .90 for subtests). Cumulative research finding also provided good sources of validity evidences of this instrument for Taiwanese children (Chen, Chang, & Yang, 2004; Chen, Lin, & Liao, 2005; Chen & Yang, 2000; Hung, Chen, & Chen, 2003; Wechsler, 1997).

Analysis of the data

Tests for the factorial invariance across normal and clinical groups were based on the analysis of covariance structure models using LISREL 8.8 (Jöreskog & Sörbom, 2006). Fig. 1 shows the structure of the hypothesized baseline model, in which 12 core WISC-III subtests (with the exception of the optional Mazes subtest)

are divided according to index-type: (1) Verbal Comprehension Index, VCI; (2) Perceptual Organizational Index, POI; (3) Freedom from Distractibility Index, FDI; and (4) Processing Speed Index, PSI. This baseline model was first tested separately so that each group could examine its appropriateness.

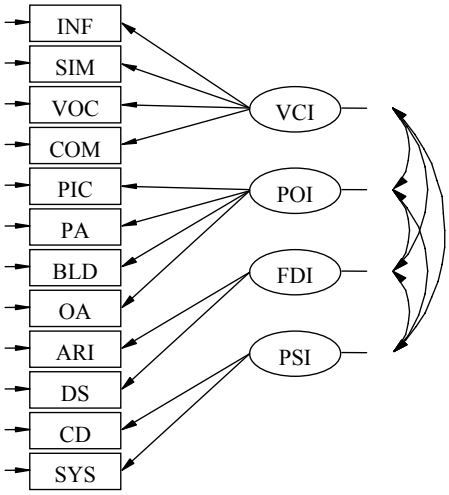

Figure 1.　Construct of the hypothesized WISC-III four-factor model

Next, invariance analyses with nested models were tested on three levels. Each level sets more constraints than are imposed by the previous one (Meredith, 1993). The first and weakest level tests for configural invariance, which assumes that the overall factor pattern is the same for both normal and clinical children. The second level tests for weak factorial invariance, or also called metric invariance; this constrained model requires the magnitude of the factor loadings be the same across all groups ($\Lambda_N = \Lambda_C$). The third level tests for uniqueness invariance ($\Theta_N = \Theta_C$); this means examining whether WISC-III four-factor-structure explains the same amount of variances of subtests in both groups. That is, whether the subtest abilities could be measured with similar accuracy across groups.

All models were tested using covariance matrices. Maximum likelihood was the estimation method because of its robustness and sensitivity to incorrectly specified models (Hu & Bentler, 1998). The scale of latent factors was defined by fixing the first factor loading as one per factor. Criteria were evaluated jointly to assess overall model fit (Bentler & Bonett, 1980; Marsh, Balla, & McDonald, 1988). These included weighted least squares χ^2, χ^2 to df ratio, goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), normed fit index (NFI), non-normed fit index (NNFI), comparative fit index (CFI), and root-mean-square-error of approximation (RMSEA). In accordance with convention, a value of .90 served as the rule-of thumb lower limit of acceptable fit for all fit indices ranging from zero to 1, with 1 indicating a perfect fit (Hoyle & Panter, 1995; Kline, 2005).

Values close to 2.0 or 3.0 were considered fit for the χ^2 to df ratio (Bollen, 1989). An RMSEA less than .05 corresponded to a "good" fit and with .08 considered an "acceptable" fit (McDonald & Ho, 2002).

During each step of the analyses, the chi square difference ($\Delta\chi^2$) was tested between nested models, and suggestions regarding to partial measurement invariance (Byrne, Shavelson, & Muthen, 1989; Byrne & Watkins, 2003) was carefully considered and followed. If inadequate fit was detected, fit in the model was improved by including additional parameters identified by the modification index (MI) provided by LISREL. Meanwhile, re-parameterization was examined carefully for meaningfulness.

Results

Descriptive statistics

The descriptive statistics for each subtest by group are presented in Table 1, along with the Shapiro-Wilk index (Shapiro & Wilk, 1965) for normality testing.

According to the mean values in Table 1, disabled children performed lower on all subtests than their normal counterparts. The pattern of the data for both groups approximated a normal distribution. Skewness ranged from -.34 to -.05 for the normal, and from .07 to .55 for the disabled group; kurtosis ranged from -.09 to .44 for the former, and from -1.08 to -.22 for the latter group. A majority of the Shapiro-Wilk indices were close to 1. Muthén and Kaplan (1985) once suggested that a likelihood estimate for variables with a skewness and kurtosis around -1 to +1 is acceptable. Kline (2005) also suggested that the z test may not be very useful in large samples because slight departures may end up as being statistically significant. An alternative is to interpret the absolute values of standardized indices. When the absolute value of skewness is larger than 3, or the absolute value of kurtosis is larger than 10, then it is considered to be a non-normal problem. Our

Table 1. Descriptive statistics for both groups

	Normal group (N = 1,100)					Clinical group (N = 1,150)				
Subtest	M	SD	Sk	Ku	S-W	M	SD	Sk	Ku	S-W
INF	10.28	3.05	-0.05	0.02	0.99	6.25	4.05	0.44	-0.46	0.94
SIM	9.74	3.65	-0.34	-0.08	0.98	6.59	4.50	0.26	-1.08	0.92
VOC	9.86	3.53	-0.31	0.07	0.98	6.53	4.03	0.28	-0.67	0.95
COM	10.21	3.30	-0.32	0.17	0.98	6.54	4.07	0.28	-0.69	0.95
PIC	10.35	3.05	-0.14	0.20	0.99	7.66	4.57	0.07	-0.97	0.95
PA	9.92	3.36	-0.24	-0.05	0.99	6.95	4.32	0.27	-0.93	0.95
BLD	10.16	3.14	-0.17	0.17	0.99	7.33	4.57	0.17	-0.89	0.95
OA	10.26	3.13	-0.19	-0.09	0.99	7.91	4.39	0.08	-0.77	0.97
ARI	10.56	3.06	-0.05	0.22	0.99	5.87	3.87	0.51	-0.22	0.94
DS	10.33	3.16	-0.06	0.06	0.99	6.72	3.81	0.33	-0.36	0.97
CD	10.14	3.17	-0.07	0.32	0.99	5.34	3.72	0.55	-0.46	0.92
SYS	10.33	3.25	-0.13	0.44	0.98	6.44	4.00	0.27	-0.62	0.95

Note: M = Mean; SD = Standard Deviation; Sk = Skewness; Ku = Kurtosis; S-W = Shapiro-Wilk test for normality

results revealed that no serious non-normality was identified in the current data, and thus the maximum likelihood method was applied for model estimation.

Baseline model checking

As indicated by all goodness-of-fit indices reported in Table 2, the initially hypothesized

Table 2. Multi-sample CFA goodness-of-fit index:

Models	χ^2	df	χ^2/ df	GFI	AGFI	NFI	NNFI	CFI	RMSEA
Phase I : Baseline model fit for each group									
Normal group (N)	176.32	48	3.67	0.97	0.96	0.99	0.99	0.99	0.049
$\Theta_{7,8}$ free	144.96	47	3.08	0.98	0.96	0.99	0.99	0.99	0.044
Clinical group (C)	449.86	48	9.37	0.94	0.90	0.98	0.98	0.99	0.085
$\Theta_{7,8}$ free	360.18	47	7.66	0.95	0.92	0.99	0.98	0.99	0.076
$\Theta_{10,12}$ free	317.23	46	6.89	0.96	0.93	0.99	0.99	0.99	0.072
$\Theta_{7,9}$ free	266.65	45	5.93	0.96	0.94	0.99	0.99	0.99	0.065
$\Theta_{1,8}$ free	236.41	44	5.37	0.97	0.94	0.99	0.99	0.99	0.062
$\Theta_{1,9}$ free	211.35	43	4.92	0.97	0.95	0.99	0.99	0.99	0.058
$\Theta_{5,8}$ free	189.04	42	4.50	0.97	0.95	0.99	0.99	1.00	0.055
$\Theta_{3,4}$ free	163.41	41	3.99	0.98	0.96	0.99	0.99	1.00	0.051
Phase II : Factor invariance across groups									
1. Configural Invariance	308.37	88	3.50	0.98	-------	0.99	0.99	0.99	0.047
2. Factor loading Invariance	338.76	96	3.53	0.98	-------	0.99	0.99	0.99	0.047
Δ2 vs.1	30.39**	8							
2a. $\Theta_{3,4}$ free (N)	318.59	95	3.35	0.98	-------	0.99	0.99	0.99	0.046
Δ2a vs. 1	10.22	7							
3. Error variance Invariance	465.33	107	4.35	0.97	-------	0.99	0.99	0.99	0.055
Δ3 vs. 2a	146.74**	12							
3a. $\Theta_{6,6}$ free (C)	440.15	106	4.15	0.97	-------	0.99	0.99	0.99	0.053
Δ3a vs. 2a	121.56**	11							
3b. $\Theta_{1,1}$ free (C)	405.31	105	3.86	0.97	-------	0.99	0.99	0.99	0.050
Δ3b vs. 2a	86.72**	10							
3c. $\Theta_{6,7}$ free (C)	383.77	104	3.69	0.97	-------	0.99	0.99	0.99	0.049
Δ3c vs. 2a	65.18**	9							
3d. $\Theta_{8,8}$ free (C)	364.09	103	3.53	0.97	-------	0.99	0.99	0.99	0.047
Δ3d vs. 2a	45.50**	8							
3e. $\Theta_{4,7}$ free (C)	342.72	102	3.36	0.98	-------	0.99	0.99	0.99	0.046
Δ3e vs. 2a	24.13**	7							
3f. $\Theta_{12,12}$ free (C)	325.93	101	3.22	0.98	-------	0.99	0.99	0.99	0.045
Δ3f vs. 2a	7. 34	6							

Note: ** p<.01; $\Theta_{7,8}$ = error covariance between Block Design and Object Assembly; $\Theta_{10,12}$ = error covariance between Digit Span and Symbol Search; $\Theta_{7,9}$ = error covariance between Block Design and Arithmetic; $\Theta_{1,8}$ = error covariance between Information and Object Assembly; $\Theta_{1,9}$ = error covariance between Information and Arithmetic; $\Theta_{5,8}$ = error covariance between Picture Completion and Object Assembly; $\Theta_{3,4}$ = error covariance between Vocabulary and Comprehension; $\Theta_{6,7}$ = error covariance between Picture Arrangement and Block Design; $\Theta_{4,7}$ = error covariance between Comprehension and Block Design; $\Theta_{6,6}$ = error variance for Picture Arrangement; $\Theta_{1,1}$ = error variance for Information; $\Theta_{8,8}$ = error variance for Object Assembly; $\Theta_{12,12}$ = error variance for Symbol Search.

four-factor model fits comparatively better for the normal group than for the clinical sample. This was a reasonable finding since Wechsler factor structures have traditionally been established based mainly on the normal population, and clinical population is known to have some distinct cognitive pattern.

We further examined the model fitness for each population individually. For the normal group, all "fit" values in this initial model were within ideal ranges. Result revealed that the four-factor structure is an appropriate construct for normal population. To identify any modification which may further improve this model, we then proceeded in an exploratory fashion to locate possible mis-fit parameters. The highest MI indicated an error covariance between Block Design and Object Assembly subtests, which was considered as a reasonable one. Once we relaxed this error covariance for the normal group, a relatively small but significant value was estimated (standardized estimate$\Theta_{7,8}$ = .13, t= 5.51, p<.01). This revised model fit was improved and thus was treated as the starting model for normal group in the following invariance checking.

The same inspection procedure was applied to the clinical sample. With an RMSEA of .085 and the χ^2 to df ratio as 9.37 shown in the initial model, we believed that some model modifications are needed for better improvement. After free estimation of seven correlated residuals being allowed, a better-fitting model was established. Since these trivial correlated residuals were considered reasonable, this improved structure was set as the starting model for the clinical sample.

Multi-sample invariance analysis

Based on the defined starting models, multi-sample analyses were conducted with constraints embedded in sequence. First, checking for configural invariance across groups (Model 1) revealed a good model-data fit. Normal and clinical children basically share the same latent four-factor structure, and corresponding subtests employ the same factors. Second, factor loadings were then constrained to be equal across groups (model 2). CFA results indicated a good model fit. However, the χ^2 difference between this model and Model 1 was significant ($\Delta\chi^2(8)=30.39$, $p<.01$), an exploratory approach was then proceeded to locate misfit parameters. The highest MI indicated an error covariance for normal group between the Vocabulary and Comprehension subtests which was considered as with appropriate meaning. Once relaxing this parameter (model 2a), a relatively small but significant value was estimated (standardized estimate $\Theta_{3,4}$ = 0.05, t= 4.30, p<.01). The revised model had improved fit and the χ^2 difference between model 2a and model 1 was not significant ($\Delta\chi^2(7)=10.22$, $p>.05$). Finally, further constraints on error variance equivalence were imposed (model 3). The model fit well from a practical perspective but with a significant χ^2 difference ($\Delta\chi^2(12)=146.74$, $p<.01$), suggesting that unique variances are not completely invariant under current model specification. MI checking again helped to indicate mis-fit parameters. With six more error variances and covariances being examined and set free for estimation, the final model (model 3f) fit the data fairly well and shown a non-significant χ^2 difference ($\Delta\chi^2(6)=7.34$, $p>.05$) compared to model 2a.

Standardized estimates based on model 3f for each group are shown in Table 3. Again, the scale of latent factors was defined by fixing the initial first factor loading estimation as one per factor. According to Table 3, while with the four-factor structure and all factor loadings remain invariant, the differences between groups mainly reside in trivial residual variance and covariance terms. In general, only a few discrepancies were identified across groups. The partial factorial invariance in factor pattern, factor loadings, and error variances between normal and clinical groups thus was supported.

Table 3. Standardized parameter estimates for invariance model 3f

	Normal group (Clinical group)												
	Factor loadings Λ	Uniqueness Θ	Residual covariances Θ										
			INF	SIM	VOC	COM	PIC	PA	BLD	OA	ARI	DS	CD	SYS
INF	.84 (.84)	.23 (.35)	- -											
SIM	.89 (.89)	.22 (.22)	0 (0)	- -										
VOC	.86 (.86)	.27 (.27)	0 (0)	0 (0)	- -									
COM	.81 (.81)	.34 (.34)	0 (0)	0 (0)	.06 (.04)	- -								
PIC	.79 (.79)	.38 (.38)	0 (0)	0 (0)	0 (0)	0 (0)	- -							
PA	.81 (.81)	.45 (.25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	- -						
BLD	.81 (.81)	.35 (.35)	0 (0)	0 (0)	0 (0)	0 (-.04)	0 (0)	0 (-.07)	- -					
OA	.72 (.72)	.40 (.56)	0 (-.06)	0 (0)	0 (0)	0 (0)	0 (.09)	0 (0)	.12 (.15)	- -				
ARI	.85 (.85)	.27 (.27)	0 (.07)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (.06)	0 (0)	- -			
DS	.70 (.70)	.52 (.52)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	- -		
CD	.69 (.69)	.53 (.53)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	- -	
SYS	.82 (.82)	.24 (.39)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (.11)	0 (0)	- -

To sum up, the normal and clinical groups differ with regard to the following trivial but statistically significant parameters: (1) Two error covariances were found resided in both normal and clinical groups, the first was the correlated residual between Block Design and Object Assembly subtests ($\Theta_{7,8=}$ .12 and .15 for each group respectively); the other was the correlated residual between Vocabulary and Comprehension subtests ($\Theta_{3,4=}$ .06 and .04 respectively); (2)

seven residual covariances were identified for the clinical group only. There were correlated errors between Digit Span and Symbol Search subtests($\Theta_{10,12}$ = .11, t = 6.10), between Block Design and Arithmetic subtests ($\Theta_{7,9}$ = .06, t = 5.31), between Information and Object Assembly subtests ($\Theta_{1,8}$ = -.06, t = -4.93), between Information and Arithmetic subtests ($\Theta_{1,9}$ = .07, t = 5.31), between Picture Completion and Object Assembly subtests ($\Theta_{5,8}$ = .09, t = 5.93), between Picture Arrangement and Block Design subtests ($\Theta_{6,7}$ = -.07, t = -5.29), also between Comprehension and Block Design subtests ($\Theta_{4,7}$ = -.04, t = -3.87); and (3) four subtests shown variant error variances across normal and clinical groups. They were Picture Arrangement ($\Theta_{6,6}$ = .45 and .25 respectively); Information($\Theta_{1,1}$ = .23 and .35 respectively); Object Assembly ($\Theta_{8,8}$ = .40 and .56 respectively); and Symbol Search($\Theta_{12,12}$ = .24 and .39 respectively).

Discussion

Findings of this study support the four-factor structure, which is consistent with many previous foreign analytic studies of the WISC-III. Also confirmed was a generally invariant factor structure across normal and clinical groups. With some exceptions regarding error variances and correlated residuals, empirical evidence generally supported that we were measuring the same theoretical latent constructs for both normal and clinical children. Besides, measurement accuracy of most WISC-III subtests was shown invariant across groups. Our findings supported partial factor invariance. In general, the WISC-III scores for both normal and clinical children could be interpreted equivalently, that is, as having the same meaning.

Byrne et al. (1989) suggested that allowing for correlated errors is often necessary in order to obtain a well-fitting model. These correlated errors usually represent nonrandom measurement errors due to method effects such as item format. They further specified that, “the equality of error variances and covariances is probably the *least* important hypothesis to test ... it is widely accepted that to do so represents an overly restrictive test of the data” (Byrne, 1998, p.261).

In this study, we found correlated residuals for Block Design and Object Assembly subtests, also for Vocabulary and Comprehension subtests in both normal and clinical samples. Item formats for these two pairs are known to be similar, for example, BLD and OA both involving a part-whole integration, VOC and COM both requiring more verbal oral expression, which are not shared by other Wechsler subtests (Sattler, 2001), these residual covariances were considered meaningful and thus could be incorporated reasonably. Nonetheless, it should be noted that the degree of group discrepancy on estimated values were fairly small for either pair (.12 vs. .15 for the former one, and .06 vs. .04 for the later one), though this might indicate a differential item format effect across groups, there seemed to be no real need for serious concern about group discrepancy.

As shown in the results, seven other subtly correlated residuals were detected in the clinical group only. These estimated values range from -.04 to .11, which were indeed quite trivial. The ones with the most significant cross-group discrepancies were the correlated residual between

Digit Span and Symbol Search, and the correlated error between Picture Completion and Object Assembly. Further data examination revealed that correlation between subtests in each pair was higher in the clinical sample (r=.55 and .71 respectively, both p<.01) then it was in the normal group (r=.34 and .45 respectively, both p<.01). This result showed that more subtest variances are actually shared in the clinical population, and it was in accordance with the known fact that cognitive abilities are less correlated within the higher ability group (Detterman, 1993; Detterman & Daniel, 1989; Legree, Pifer, & Grafton, 1996; Lynn, 1990; Lynn & Cooper, 1993, 1994; Spearman, 1927). Sweetland, Reina, and Tatti (2006) suggested that lower ability children may have deficits in important central cognitive processes, and thus tend to operate on a more uniform lower level.

Further more, it is clear from a review of the literature that Digit Span and Symbol Search both require children to form visual images, utilize visual scanning (either physically or mentally), have a tolerance for stress, and be able to concentrate; meanwhile, Picture Completion and Object Assembly both tap children's ability to form wholistic concepts on pictures with meaningful contents (Sattler & Dumont, 2004). All those correlated error terms in the clinical sample might represent some shared components which are more congruent in the clinical population. It might also be reasonable to assume that, for clinical children, once partitioning off the shared variances between subtests which could be explained by latent factors identified in this four-factor model, some trivial portion could remain un-explained. However, given the subtlety of such differences across groups, we believe that the overall main structure for WISC-III can be fairly considered as stable enough across normal and clinical populations.

Finally, four subtests were found with discrepant error variances across groups. The error variances for Information, Object Assembly, and Symbol Search subtests were estimated slightly higher in the clinical group, which revealed that the corresponding tested intellectual abilities could be explained more thoroughly by the four-factor model in the normal children population. On the contrary, the error variances for Picture Arrangement subtest was estimated slightly higher in the normal group, thus suggested that this kind of ability could be better explained by the model in the clinical children population. Many factors, such as population unique variation or ability level, could contribute to these discrepant findings, or it could be the interaction between test contents and population exceptionality. While we remind researchers to be careful when explaining the test result for children in different population, the current findings surely deserve further examinations.

Since our primary goal was to get a wholistic view on factorial invariance across groups, some inevitable limitations of the present study deserve attention. First, even if the partial measurement invariance approach helped us getting the maximum information regarding the degree of overall invariance, we realized that the exploratory post-hoc approach in identifying possible variant parameters could bring the risk of identification on chance level. In this study, besides the statistical MI indicators, we did monitor each step and try to reason each modification

with sensible meaning through the whole model fitting process. Nonetheless, as Vandenberg and Lance (2000) suggested, researchers must carefully consider and examine the theoretical justification when relaxing constraints in practicing the partial invariance study. Second, it is good that we used large and heterogeneous samples for ensuring the stability and accuracy for parameter estimations in this current study. However, with such detailed model-fitting modifications, we realize that it could be a problem for later cross-validation. Researchers are encouraged to aware of this limitation.

In summary, given the relatively large data set and substantial number of variations with which we were working, the findings were of substantial importance for understanding the factorial structure for this frequently used instrument. Consistent with contemporary research findings, the underlying WISC-III factor structure is appropriately represented by the proposed four-factor model comprising VCI, POI, FDI, and PSI. Moreover, except for some trivial discrepancies on error variances and correlated residuals, majorities of the model parameters were demonstrated invariant across groups. Evidence of multi-level invariance supported the partial factor invariance of this instrument across normal and clinical groups. The main structure and the underlying meaning of each WISC-III factor are generally identical for both normal and clinical children populations in Taiwan.

References

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

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

Bollen, K. A.(1989). *Structural equations with latent variables*. New York:Wiley.

Byrne, B. M. (1998). *Structure equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of cross-cultural psychology, 34*(2), 155-175.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,* 141-154.

Chen, H., Chang, C., & Yang, T. (2004). Study of intellectual patterns of children with Autism. *Bulletin of Special Education*, *26*, 127-151.

Chen, H., Lin, K., & Liao, Y. (2005). Study of WISC-III Intellectual Patterns of Mental Retarded Children. *Bulletin of Special Education*, *28*, 97-122.

Chen, H., & Yang, T. (2000). Base Rates of WISC-III diagnostic subtest patterns in Taiwan: Standardization , Learning Disabled

and ADHD samples applied. *Psychological Testing*, *47*(2), 91-110.

Chen, H., Zhu, J., & Chen, Y. (2000). *The legitimacy and utility of the WISC-III factor-based indexes: Taiwan standardization sample applied*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, U.S.A.

Detterman, D. K. (1993). Giftedness and intelligence: One and the same? In Ciba Foundation (Ed.), *The origins and development of high ability*. West Sussex, England: Wiley.

Detterman, D. K., & Daniel, M. H. (1989).Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, *13*, 349-359.

Donders, J., & Warschausky, S. (1996). A structural equation analysis of the WISC-III in children with traumatic head injury. *Child Neuropsychology, 2*, 185-192.

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

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

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

Georgas,J., Van de Vijver, F. J. R., Weiss, L. G., & Saklofske, D. H. (2003). A cross-cultural analysis of the WISC-III. In J. Georgas, L. G. Weiss, F. J. R. Van de Vijver, & D. H. Saklofske (Eds.), *Culture and children's intelligence: Cross-cultural analysis of the WISC-III* (pp. 277-313). San Diego, CA: Academic Press.

Horn, J. L., & McArdle. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117-144.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: concepts, issues, and applications*. 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.

Hung, L., Chen, S., & Chen, H. (2003). The Study of Intelligence of Secondary Students with Learning Disabilities. *Journal of Taiwan Normal University: Education*, *48*(2), 215-238.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 statistical programs*. Chicago: Scientific Software.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. NY: John Wiley & Sons.

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

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). NY: The Guilford Press.

Konold, T. R., Kush, J. C., & Canivez, G. L.

(1997). Factor replication of the WISC-III in three independent samples of children receiving special education. *Journal of Psychoeducational Assessment, 15*, 123-137.

Legree, P. J., Pifer, M. E., & Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher ability groups. *Intelligence, 23*, 45-57.

Lynn, R. (1990). Does Spearman's g decline at high levels? Some evidence from Scoland. *The Journal of Genetic Psychology*, *153*, 229-230.

Lynn, R., & Cooper, C. (1993). A secular decline in Spearman's g in France. *Learning and Individual Differences*, *5*, 43-48.

Lynn, R., & Cooper, C. (1994). A secular decline in Spearman's g in Japan. *Current Psychology Developmental, Learning, Personality, Social, 13*, 3-9.

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*, 391-410.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equating analyses. *Psychological Methods, 7*(1), 64-82.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E., & Kwok, O.(2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

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

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13*, 403-418.

Roid, G. H., Prifitera, A., & Weiss, L. G. (1993). Replication of the WISC-III factor structure in an independent sample. *Journal of Psychoeducational Assessment Monograph Series. Advances in Psychological Assessment: Wechsler Intelligence Scale for Children-Third Edition*, 6-21.

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

Sattler, J. M. (2001). *Assessment of children cognitive applications* (4th ed). Le Mesa, CA: Author.

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

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*, 591-611.

Spearman, C. (1927). *The ability of man*. London: MacMillan.

Sweetland, J. D., Reina, J. M., & Tatti, A. F. (2006). WISC-III verbal/performance dis-

crepancies among a sample of gifted children. *Gifted Child Quarterly, 50*(1), 7-10.

Tupa, D. J., Wright, M. O., & Fristad, M. A. (1997). Confirmatory factor analysis of the WISC-III with child psychiatric inpatients. *Psychological Assessment, 9*(3), 302-306.

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.

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

Wechsler, D. (1997). *Manual for the Taiwan version of the Wechsler Intelligence Scale for Children - Third Edition.* Taipei, Taiwan: Chinese behavioral science corporation.

Weiss, L. G., Saklofske, D. H., Schwartz, D. M., Prifitera, A., & Courville, T. (2006). Advanced clinical interpretation of WISC-IV index scores. In. L.G. Weiss, D. H. Saklofske, A. Prifitera, & J. A. Holdnack. (Eds.), *WISC-IV advanced clinical interpretation* (pp139-179). Burlington, MA: Academic Press.

收稿日期：2007.10.12
接受日期：2008.03.11

Bulletin of Special Education
2008, 33(1), 93-107

臺灣一般兒童與臨床兒童之魏氏兒童智力量表第三版(WISC-III)因素恆等性研究

陳心怡
臺灣師範大學特教系教授

朱建軍
美國心理公司主任

洪儷瑜
臺灣師範大學特教系教授

張正芬
臺灣師範大學特教系教授

楊宗仁
臺北教育大學特教系副教授

本研究主要目的在檢驗魏氏兒童智力量表第三版(WISC-III)在一般兒童與特殊兒童組群間之因素恆等性。文中根據 1,100 名 WISC-III 標準化樣本及 1150 名臨床兒童樣本資料,運用結構方程模式進行多樣本驗證性因素分析。本研究以 WISC-III 12 個主要分測驗與四個潛在因素架構為假設基準模型，分二階段完成檢驗。階段一先分就二組進行單樣本驗證性因素分析，以建立適當之基準因素模式。階段二則建立階段性逐步設限的巢套模型，逐一檢驗因素架構、因素負荷量、與殘差變異數之跨組別恆等假設。研究發現除了少處細微之殘差變異或共變性差異外,絕大多數之檢驗參數均具跨樣本恆等特性，WISC-III 在一般兒童與特殊兒童組群間具有相同之四因素架構與因素負荷量。整體而言，研究結果支持部份因素恆等性，台灣臨床兒童與一般兒童之 WISC-III 分數是具有相同之解釋意義。

關鍵詞：特殊兒童、因素恆等性、多樣本驗證性因素分析、WISC-III