School Psychology Quarterly
2014, Vol. 29, No. 1, 52–63

© 2013 American Psychological Association
1045-3830/14/$12.00 DOI: 10.1037/spq0000038

Bifactor Structure of the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition

Marley W. Watkins and A. Alexander Beaujean
Baylor University

The Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition (WPPSI-IV; Wechsler, 2012) represents a substantial departure from its predecessor, including omission of 4 subtests, addition of 5 new subtests, and modification of the contents of the 5 retained subtests. Wechsler (2012) explicitly assumed a higher-order structure with general intelligence (*g*) as the second-order factor that explained all the covariation of several first-order factors but failed to consider a bifactor model. The WPPSI-IV normative sample contains 1,700 children aged 2 years and 6 months through 7 years and 7 months, bifurcated into 2 age groups: 2:6–3:11 year olds ($n = 600$) and 4:0–7:7 year olds ($n = 1{,}100$). This study applied confirmatory factor analysis to the WPPSI-IV normative sample data to test the fit of a bifactor model and to determine the reliability of the resulting factors. The bifactor model fit the WPPSI-IV normative sample data as well as or better than the higher-order models favored by Wechsler (2012). In the bifactor model, the general factor accounted for more variance in every subtest than did its corresponding domain-specific factor and the general factor accounted for more total and common variance than all domain-specific factors combined. Further, the domain-specific factors exhibited poor reliability independent of *g* (i.e., ω_h coefficients of .05 to .33). These results suggest that only the general intelligence dimension was sufficiently robust and precise for clinical use.

Keywords: WPPSI-IV, factor analysis, intelligence, preschool, reliability

Supplemental materials: http://dx.doi.org/10.1037/spq0000038.supp

Wechsler intelligence scales are widely used throughout the world (Flanagan & Kaufman, 2009). For example, the third edition of the Wechsler Intelligence Scale for Children (WISC–III) was adapted for at least 12 locations (Georgas, van de Vijver, Weiss, & Saklofske, 2003) and the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV; Wechsler, 2003) have been adapted and standardized in seven countries (Grégoire et al., 2008). Likewise, the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) has been adapted for use in other countries (Liu & Lynn, 2011).

The fourth edition of the WPPSI (WPPSI-IV; Wechsler, 2012), was recently completed in the United States. As with new editions of other Wechsler instruments (Coalson & Weiss, 2002), the WPPSI-IV represents a substantial departure from its predecessor, the third edition of the WPPSI (WPPSI-III; Wechsler, 2002). Specifically, the WPPSI-IV removed four subtests, added five new subtests, and modified the contents of the five retained subtests. Differences between a test and its revision are "expected and derive from differing procedures, measurement scales, and normative bases" (L. D. Nelson, 2000, p. 235). Consequently, validity evidence from prior versions of the WPPSI (e.g., Sattler, 2008) may not be applicable to the WPPSI-IV (Reise, Waller, & Comrey, 2000). Accordingly, Wechsler (2012) presented extensive validity evidence for the WPPSI-IV, including evidence on its structural validity.

Given that an instrument's structural validity evidence involves a comparison of how well its scores measure the trait(s) it intends to measure

This article was published Online First November 4, 2013.

Marley W. Watkins and A. Alexander Beaujean, Department of Educational Psychology, Baylor University.

Correspondence concerning this article should be addressed to Marley W. Watkins, Department of Educational Psychology, Baylor University, Waco, TX 76798-7301. E-mail: Marley_Watkins@baylor.edu

(American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), it is vital that alternative structural models be examined (Floyd & Widaman, 1995). For the WPPSI-IV, Wechsler (2012) explicitly assumed a higher-order factor structure with general intelligence (*g*) as the second-order factor that explained the covariation among the first-order factors. In this model, the relationship between *g* and every subtest (i.e., observed variable) is fully mediated by the first-order factors (Yung, Thissen, & McLeod, 1999) and *g* is viewed as a superordinate construct (Gignac, 2006). See Figure 1 for an example of a second-order factor structure of the WPPSI-IV subtests.

Because first-order factors are abstractions of measured variables, second-order factors "are abstractions of abstractions even more removed from the measured variables" (Thompson, 2004, p. 81). Gorsuch (1983, p. 245) argued that "basing interpretations upon interpretations" in this manner is undesirable and recommended that factors be directly related to observed variables. This can be accomplished with a bifactor model (Holzinger & Swineford, 1937), sometimes called a *nested-factors* (Gustafsson & Undheim, 1996) or *direct hierarchical* (Gignac, 2006) model, which specifies that every factor has a direct effect on the observed variables. Gignac (2006, p. 85) argued that it is "more congruent and reasonable to specifically model the most significant factor of a battery of tests (i.e., "g") directly, rather than indirectly, through first-order factors." In a typical bifactor model, each subtest is directly and independently influenced by two factors: one general factor and one domain-specific first-order factor.[1] *g* is conceptualized as a breadth factor in the bifactor model, which is consistent with Spearman's (1927) conceptualization of general intelligence (cf. Carroll, 1996; for an alternative interpretation, see Reynolds & Keith, 2013). See Figure 2 for an example of a bifactor structure of the WPPSI-IV subtests.

Reise (2012) argued that the bifactor model is a viable candidate for measures that have demonstrated good fit to a second-order model, as was found with the WPPSI-IV (Wechsler, 2012). The bifactor model has been specifically recommended for tests of a variety of constructs, such as intelligence (Brunner, Nagy, & Wilhelm, 2012), health outcomes (Reise, Morizot, & Hays, 2007), quality of life (Varni, Beaujean, & Limbers, in press), psychiatric distress (Thomas, 2012), early academic skills (Betts, Pickart, & Heistad, 2011), personality (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012), and psychopathology (Brouwer, Meijer, & Zevalkink, 2013). Although the bifactor model has been found to be a good fit for other Wechsler scales (Brunner et al., 2012; Gignac, 2006; Golay & Lecerf, 2011; Golay, Reverte, Rossier, Favez, & Lecerf, 2013; Gustafsson & Undheim, 1996; J. M. Nelson, Canivez, & Watkins, 2013; Watkins, 2010), Wechsler (2012) did not examine its applicability for the WPPSI-IV. This is an unfortunate omission because, unlike higher order models, bifactor models allow an examination of the strength of the direct relationship between the observed variables and the factors as well as a determination of the role of first-order factors independent of the general factor (Chen et al., 2006, 2012).

These benefits of bifactor models are particularly salient when examining cognitive ability data because a factor or variable that does not contribute beyond what it shares with the general factor might be identified (Chen et al., 2012), suggesting that the data is overfactored (e.g., Frazier & Youngstrom, 2007). For example, general and fluid intelligence factors were found to be redundant for the Wechsler Adult Intelligence Scale (Fourth Edition) (Wechsler, 2008) when a bifactor model was applied (Nileksela, Reynolds, & Kaufman, 2013). In addition, an observed variable that reflects only *g* may fail to emerge in the bifactor model and will likely result in estimation problems, such as small loadings on a domain-specific first-order factor (Rindskopf & Rose, 1988). This would occur because the common variance in the measured variable is entirely explained by *g*. Such problems would not be easy to detect with higher-order models (Chen et al., 2006). Keith's (2005) analysis of the WISC-IV provided an example of this situation, as the Arithmetic subtest had a loading of .80 on its first-order factor in the higher-order model but only .11 in the bifactor model (cf. Reynolds & Keith, 2013).

[1] Under certain conditions, higher-order and bifactor models can be made equivalent (Yung et al., 1999), but as their purpose and interpretation are usually quite different (Chen, West, & Sousa, 2006), we treat them as separate models.

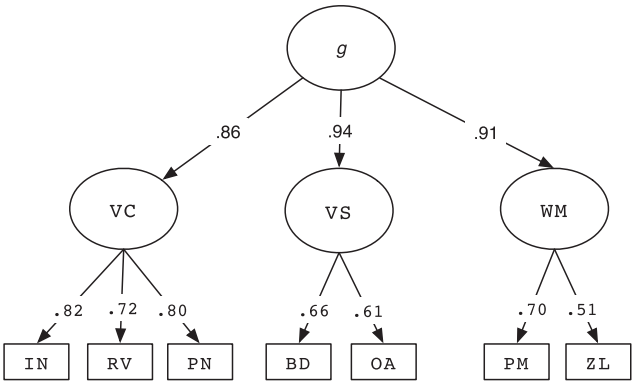

Figure 1. Example of second-order structure for the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition. VC = Verbal Comprehension factor; VS = Visual-Spatial factor; WM = Working Memory factor; IN = Information; RV = Receptive Vocabulary; PN = Picture Naming; BD = Block Design; OA = Object Assembly; PM = Picture Memory; ZL = Zoo Locations.

Given these advantages, the current study examined the applicability of a bifactor model for the WPPSI-IV normative sample data.

Method

Participants

The WPPSI-IV normative sample contains 1,700 English-speaking children aged 2 years and 6 months through 7 years and 7 months, bifurcated into two age groups: 2:6–3:11 year olds (n = 600) and 4:0–7:7 year olds (n = 1,100). The standardization sample closely matched the 2010 census on gender, age, race/ethnicity, parent education level, and geographic region. Across half-year age increments, the sample was 12.0% to 16.5% African American, 2.0% to 5.0% Asian American, 22.0% to 27.0% Hispanic, 50.0% to 57.5% White, and 3.0% to 5.5% of other race/ethnicity. Wechsler (2012, Chapter 3)

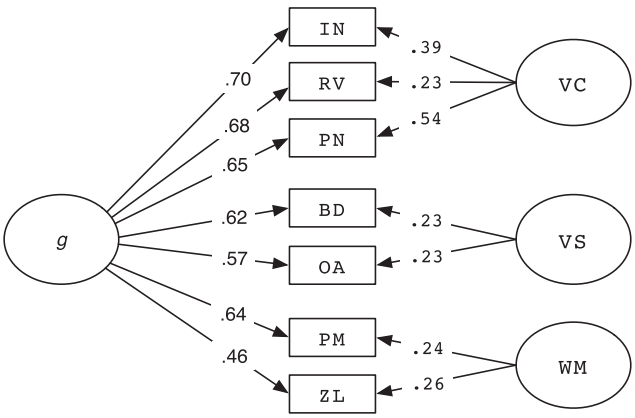

Figure 2. Bifactor model for the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition among 600 children aged 2:6 to 3:11 years. VC = Verbal Comprehension factor; VS = Visual-Spatial factor; WM = Working Memory factor; IN = Information; RV = Receptive Vocabulary; PN = Picture Naming; BD = Block Design; OA = Object Assembly; PM = Picture Memory; ZL = Zoo Locations. The domain-specific loadings for the VS and WM factors were constrained to be equal.

provided detailed information on the standardization sample grouped across the stratification variables.

Instrument

At ages 2:6 through 3:11 years, the WPPSI-IV subtests produce three Primary Index scores, each composed of two subtests: (a) the Verbal Comprehension Index (VC) from the Receptive Vocabulary and Information subtests; (b) the Visual Spatial Index (VS) from the Block Design and Object Assembly subtests; and (c) the Working Memory Index (WM) from the Picture Memory and Zoo Locations subtests. One supplemental subtest, Picture Naming, is provided.

At ages 4:0 through 7:7 years, the WPPSI-IV subtests produce five Primary Index scores, each composed of two subtests: (a) the VC from the Information and Similarities subtests; (b) the VS from the Block Design and Object Assembly subtests; (c) the Fluid Reasoning Index (FR) from the Matrix Reasoning and Picture Concepts subtests; (d) the WM from the Picture Memory and Zoo Locations subtests; and (e) the Processing Speed Index (PS) from the Bug Search and Cancellation subtests. An additional five supplemental subtests are provided.

Wechsler (2012) reported that the average internal consistency reliability of WPPSI-IV subtests ranged from .75 for Animal Coding to .93 for Similarities and the average internal consistency reliabilities of Primary Index scores ranged from .86 for PS to .94 for VC (p. 49), and these high reliability coefficients generalized across a variety of clinical samples (p. 51). Short-term stability coefficients were high as well (.75 to .87 for subtests and .84 to .89 for Indexes, pp. 56–59). Wechsler presented considerable validity evidence. For example, correlations between the WPPSI-IV FSIQ and the omnibus score from other test batteries ranged from .81 with the DAS-II to .86 with the WPPSI-III and the correlation between WPPSI-IV FSIQ and academic achievement was .75 (pp. 84–99).

Analyses

Confirmatory factor analyses. Wechsler (2012, pp. 75–83) presented a series of confirmatory factor analytic (CFA) models, ranging from a single factor to a second-order factor model for each age group. Although Wechsler (2012) provided no information about the constraints used to identify these models, we were able to reproduce the results using effects coding. The effects-coding method of identification constrains the set of indicator loadings for a given factor to average 1.0, or, equivalently, constrains the sum of a factor's loadings to be equal to the number of observed indicator variables for the factor (Little, Slegers, & Card, 2006). This method of identification makes the factor loading estimates an "optimal balance" around 1.0 but does not constrain any particular loading to be 1.0. The result is that the latent variance estimates "reflect the observed metric of the indictors, optimally weighted by the degree to which each indicator represents the underlying latent construct" (Little et al., 2006, p. 63).

We replicated the Wechsler (2012) analyses for each age group using all subtests, but also fit a bifactor model. All models were fit using the lavaan package (Rosseel, 2012) in the R programming language (R Development Core Team, 2011) using maximum likelihood estimation. Wechsler (2012, pp. 80–82) reported CFA models using all the subtests as well as models using just the subtests that comprise the Primary Index scores. We performed analyses using both sets of variables, but the results did not substantially differ. To conserve space, only the results from all the subtests are included in this paper. However, the R syntax used to fit all models as well as the results from analyses with both sets of variables are available in the online ancillary materials.

Model fit. Although there are no universally accepted cutoff values for model fit indices (West, Taylor, & Wu, 2012), we needed some criteria to judge the models. Thus, we examined multiple indices that represented a variety of fit criteria (Marsh, Hau, & Grayson, 2005). Specifically, the (a) χ^2, (b) comparative fit index (CFI), (c) root mean square error of approximation (RMSEA), (d) standardized root-mean-square residual (SRMR), and (e) Akaike's Information Criterion (AIC). For good model-data fit criteria, we used the following guidelines: (a) CFI $\geq$ 0.95; (b) RMSEA $\leq$ 0.06; and (c) SRMR $\leq$ 0.06 (Hu & Bentler, 1999; Sivo, Xitao, Witta, & Willse, 2006). There are no specific criteria for information-based fit indices like the AIC, but when comparing two models from the same

data, smaller values indicate better approximations of the true model (Markon & Krueger, 2004). For a model to be deemed superior, it had to (a) exhibit good fit according to CFI, RMSEA, and SRMR indices; (b) demonstrate a ΔCFI value ≥0.01 for nested models (Dimitrov, 2012); and/or (c) display the smallest AIC value (Burnham & Anderson, 2004).

Model-based reliability. The bifactor model hypothesizes that each WPPSI-IV subtest is independently influenced by two latent constructs: the general ability factor (*g*) and one broad domain-specific first-order factor (e.g., VC, VS, etc.). Traditional internal consistency measures (i.e., alpha) assume that all consistent variability is true score variance from a single construct and that all items are equally sensitive in measuring that construct (Yang & Green, 2011). This is likely inappropriate for scores where a bifactor model fits the data well, as the consistent variance contributed by both *g* and the domain-specific factor is inappropriately attributed to the domain-specific first-order factor alone (Yang & Green, 2011). Unbiased alternative measures of construct reliability are omega (ω) and omega hierarchical (ω_h; Zinbarg, Revelle, Yovel, & Li, 2005). ω estimates the variance accounted for by both constructs (i.e., general *and* domain-specific) influencing indicators in a given domain, whereas ω_h estimates the variance accounted for by a single target construct (i.e., either general or domain-specific first-order). Using the computational methods of Brunner et al. (2012), ω and ω_h were calculated with the omega software package (Watkins, 2013). Although there are no definitive standards for ω and ω_h (Reise, Bonifay, & Haviland, 2013), Shrout and Lane (2012, p. 646, emphasis in original) suggested, ".00 to .10, *virtually no reliability*; .11 to .40, *slight*; .41 to .60, *fair*; .61 to .80, *moderate*; .81 to 1.0; *substantial reliability*."

Results

Ages 2:6–3:11 Group

Using all seven subtests, Wechsler (2012) evaluated three models: (a) a single-factor model (*g*); (b) a second-order model with one second-order factor (*g*) and two first-order factors (Verbal and Nonverbal factors); and (c) a second-order model with one second-order factor (*g*) and three first-order factors (VC, VS, and WM). We added a bifactor version of Wechsler's (2012) third model. As displayed in

Table 1
Goodness-of-Fit Statistics for Confirmatory Factor Analyses for Age Groups 2:6–3:11 (n = *600) and 4:0–7:7* (n = *1,100)*

Model	χ^2	*df*	*p*	CFI	RMSEA	SRMR	AIC
Ages 2:6–3:11							
1-General factor	70.17	14	.01	.96	.08	.04	20095
2-General, verbal, nonverbal	31.46	10	.01	.98	.06	.02	20066
3-General, VC, VS, WM	25.46	8	.01	.99	.06	.02	20064
Bifactor version of Model 3	13.73	9	.13	1.00	.03	.01	20048
Ages 4:0–7:7							
1-General factor	950.32	90	.01	.88	.09	.06	76675
2-General, verbal, nonverbal	500.34	86	.01	.94	.07	.04	76235
3-General, VC, PS, (VS + FR + WM)	284.59	84	.01	.97	.05	.03	76023
4a-General, VC, WM, PS, (VS + FR)	270.11	82	.01	.97	.05	.03	76013
4b-General, VC, VS, PS, (FR + WM)	263.22	82	.01	.97	.04	.03	76006
5a-General, VC, VS, FR, WM, PS	249.64	80	.01	.98	.04	.03	75996
Bifactor version of Model 5a	231.47	75	.01	.98	.04	.02	75998
5b-General, VC(1), VC(2), VS, FR, WM, PS	212.03	76	.01	.98	.04	.02	75967
Bifactor version of Model 5b	191.61	74	.01	.98	.04	.02	75962

Note. Models are labeled as per Wechsler (2012, pp. 76–77). Also following Wechsler, all models were identified using effects-coding (Little, Slegers, & Card, 2006). VC = Verbal Comprehension factor; VC(1) = Verbal Comprehension first subfactor; VC(2) = Verbal Comprehension second subfactor; VS = Visual-Spatial factor; WM = Working Memory factor; PS = Processing Speed factor; FR = Fluid Reasoning factor; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root-mean-square residual; AIC =Akaike's Information Criterion.

Table 1, three of the four models exhibited good fit, but the bifactor model fit the data better (ΔCFI = .01 and ΔAIC = 16) than the higher-order model preferred by Wechsler. Thus, it was deemed the superior model.

Using the coefficients from the bifactor model, we examined the reliability of each factor (see Table 2 and Figure 2). The *g* factor accounted for more total and common variance (38.5% and 78.7%, respectively) than all domain-specific first-order factors combined. In addition, it exerted a stronger direct influence on each subtest than the corresponding domain-specific factor. The ω_h coefficient for the *g* factor was .78, but the domain-specific factors had poor reliability independent of *g* (i.e., ω_h coefficients of .08 to .20). None of the VS or WM subtests had loadings on the domain-specific factors ≥ .30, and only two of the three VC subtests exhibited domain-specific factor loadings ≥ .30. Similarly, only the three subtests (all measuring VC) exhibited communali-

Table 2
Sources of Variance in the Wechsler Preschool and Primary Scale of Intelligence (4th Ed.) Among 600 Children Aged 2:6 to 3:11 Years and 1,100 Children Aged 4:00–7:7 Years

	General		VC(1)		VC(2)		VS		WM		FR		PS			
Subtest	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	h^2	u^2
Ages 2:6–3:11 Years																
IN	.70	49.0	**.39**	15.2											64.2	35.8
RV	.68	46.2	.23	5.3											51.5	48.5
PN	.65	42.3	**.54**	29.2											71.4	28.6
BD	.62	38.4					.23	5.3							43.7	56.3
OA	.56	31.4					.23	5.3							36.7	63.3
PM	.64	41.0							.24	5.8					46.7	53.3
ZL	.46	21.2							.26	6.8					27.9	72.1
Total		38.5		7.1				1.5		1.8					48.9	51.1
Common		78.7		14.5				3.1		3.7						
ω	.85		.83				.57		.54							
ω_h	.78		.20				.08		.09							
Ages 4:0–7:7 Years																
IN	.67	44.9	**.42**	17.6											62.5	37.5
SI	.66	43.6	**.49**	24.0											67.6	32.4
VO	.65	42.3	**.47**	22.1											64.3	35.7
CO	.61	37.2	**.48**	23.0											60.3	39.7
RV	.63	39.7			**.36**	13.0									52.6	47.4
PN	.61	37.2			**.57**	32.5									69.7	30.3
BD	.67	44.9					**.30**	9.0							53.9	46.1
OA	.63	39.7					.29	8.4							48.1	51.9
MR	.69	47.6									.18	3.2			50.8	49.2
PC	.59	34.8									.18	3.2			38.0	62.0
PM	.62	38.4							.28	7.8					46.3	53.7
ZL	.54	29.2							.27	7.3					36.5	63.5
BS	.58	33.6											**.49**	24.0	57.7	42.3
CA	.45	20.3											**.40**	16.0	36.3	63.7
AC	.51	26.0											**.50**	25.0	51.0	49.0
Total		37.3		5.8		3.0		1.2		1.0		0.4		4.3	53.0	47.0
Common		70.3		10.9		5.7		2.2		1.9		0.8		8.2		
ω	.93		.88		.76		.68		.58		.61		.74			
ω_h	.86		.30		.27		.12		.11		.05		.33			

Note. *b* = standardized factor loading; Var = % variance explained; h^2 = communality; u^2 = uniqueness; VC(1) = Verbal Comprehension factor or Verbal Comprehension first subfactor; VC(2) = Verbal Comprehension second subfactor; VS = Visual-Spatial factor; WM = Working Memory factor; PS = Processing Speed factor; FR = Fluid Reasoning factor; IN = Information; RV = Receptive Vocabulary; PN = Picture Naming; BD = Block Design; OA = Object Assembly; PM = Picture Memory; ZL = Zoo Locations; SI = Similarities; VO = Vocabulary; CO = Comprehension; MR = Matrix Reasoning; PC = Picture Concepts; BS = Bug Search; CA = Cancellation; AC = Animal Coding; ω = omega; and ω_h = omega hierarchical. Factor loadings ≥ .30 are in bold.

ties larger than their uniqueness, indicating that much of the variability in these subtests is comprised of subtest-specific and error variance.

Ages 4:0–7:7 Group

Using all 15 subtests at the 4:0 through 7:7 year group, Wechsler (2012, pp. 76–80) examined seven models, labeled Models 1, 2, 3, 4a, 4b, 5a, and 5b, with their preferred models (5a and 5b) being higher-order models with five first-order factors. Model 5b differed from 5a in that it added two VC "subfactors": (a) Broad/Expressive, comprised of four subtests, and (b) Focused/Simple, comprised of two subtests. We estimated bifactor versions of both five-factor higher-order models. As illustrated in Table 1, seven of the nine models displayed good fit to the data, but the bifactor models fit better than their corresponding higher-order counterparts (i.e., $\Delta AIC = 5$ for bifactor Model 5b vs. Wechsler Model 5b). Consequently, the bifactor model was deemed superior.

Using the estimates from a bifactor version of Model 5b, the *g* factor accounted for more total and common variance (37.3% and 70.3%, respectively) than all domain-specific first-order factors combined—a result replicated in both age groups. In addition, *g* exerted a stronger direct influence on each subtest than its corresponding domain-specific factor (see Table 2 and Figure 3). The ω_h coefficient for the *g* factor was .86, but the domain-specific factors had

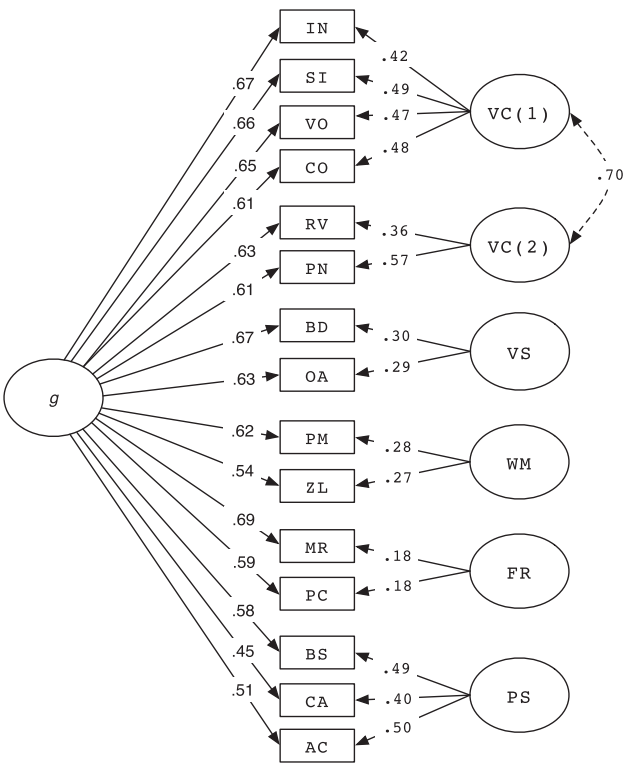

Figure 3. Bifactor model for the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition among 1,100 children aged 4:00–7:7 years. VC(1) = Verbal Comprehension factor or Verbal Comprehension first subfactor; VC(2) = Verbal Comprehension second subfactor; VS = Visual-Spatial factor; WM = Working Memory factor; FR = Fluid Reasoning factor; PS = Processing Speed factor; IN = Information; SI = Similarities; VO = Vocabulary; CO = Comprehension; RV = Receptive Vocabulary; PN = Picture Naming; BD = Block Design; OA = Object Assembly; PM = Picture Memory; ZL = Zoo Locations; MR = Matrix Reasoning; PC = Picture Concepts; BS = Bug Search; CA = Cancellation; AC = Animal Coding.

poor reliability independent of *g* (i.e., ω_h coefficients of .05 to .33). The VC, VS, and PS factors all had subtests that exhibited factor loadings on the domain-specific factors ≥.30, but none of the subtests for the WM or FR factors had factor loadings ≥.30. Similarly, most of the WM and FR subtests exhibited communalities smaller than their uniqueness, indicating that subtest-specific and error variance predominated those four subtests' variance.

Discussion

A bifactor model fit the WPPSI-IV normative sample data as well as or better than the higher-order models favored by Wechsler (2012). Unlike higher-order models, a bifactor model allows direct examination of the strength of the relationship between WPPSI-IV subtests and their corresponding domain-specific factors. As quantified in Table 2, the general factor accounted for more variance in every subtest than did its corresponding domain-specific first-order factor. In addition, the *g* factor accounted for more total and common variance than all domain-specific first-order factors combined.

Bifactor models allow for an analysis of the role of the domain-specific first-order factors independent of the general factor. As reflected in Table 2, the strongest WPPSI-IV domain-specific first-order factor (VC) accounted for between 15%–17% of the common variance. In addition, many subtests exhibited communalities smaller than their uniqueness, indicating that much of the observed variability on those subtests is due to either unique aspects of those subtests or measurement error. Further, the domain-specific first-order factors exhibited poor reliability independent of *g* (i.e., ω_h coefficients of .05 to .33). These results suggest that general intelligence should be conceptualized as a first-order breadth factor rather than a higher-order superordinate factor.

Interpretation of scores from tests like the WPPSI-IV is often ambiguous because of their multidimensional structure. A total score (e.g., FSIQ) will be more reliable than any subscore (e.g., VCI, WMI, etc.), but may be confounded by systematic variance from subscore constructs. Likewise, subscores can be confounded by shared systematic variance and therefore may not provide unique information beyond that contributed by the total score (Chen et al., 2012). Clarification of this ambiguity can be achieved if the proportion of variance in observed scores due to a single common latent variable can be determined. One way to accomplish this variance partitioning is by estimating a bifactor structure and computing model-based reliability indices (Reise et al., 2013). For the WPPSI-IV total and subscores, ω estimates the proportion of variance that can be attributed to all sources of common variance and ω_h estimates the proportion of variance that can be attributed to a single common latent variable. Table 2 reveals that ω and ω_h were relatively equivalent for the general intelligence factor ($\Delta = .07$), but highly dissimilar for the domain-specific factors (Δ.41 to Δ.63). Thus, the total score can be interpreted as substantially reflecting the intended construct of general intelligence. In contrast, the subscores contained large proportions of common and unique variance and consequently cannot be interpreted as pure measures of the single latent variables they purport to measure (e.g., verbal comprehension, working memory, etc.).

Accordingly, this study's results mitigate against Wechsler's (2012, p. 144) reliance on the domain-specific index scores "as the principal level of clinical interpretation." To the contrary, these results suggest that only the general intelligence dimension (and its manifestation in the full scale IQ score) is sufficiently robust and reliable for clinical use of such a high stakes instrument (Bergeron & Floyd, 2013). Further, the poor reliability of domain-specific first-order factors, low communality of subtests, and weak loadings of subtests on domain-specific first-order factors might be symptoms of overfactoring (Frazier & Youngstrom, 2007).

In the absence of contradictory evidence, practitioners should (a) interpret the factor structure of the WPPSI-IV as consisting of a general intelligence factor and the several primary index factors (depending on age) described by Wechsler (2012); (b) afford predominant interpretive weight to the FSIQ (as a proxy for general intelligence) because it captured the greatest amount of common variance and was the most reliable construct; (c) interpret the domain-specific index scores with great caution, remembering that their poor reliability and the high levels of subtest-specific and error variance found in their constituent subtests will not allow them to "necessarily provide additional and

separate information" (Golay et al., 2013, p. 507); and (d) remain alert for new validity evidence that might enhance these interpretive suggestions.

One possible limitation of this study is the use of effects-coding for factor identification. When Little et al. (2006) introduced the method, they did so in the context of multigroup models and only suggested its use for first-order models with simple structure. As this was the only method able to reproduce Wechsler's (2012) results, we used it for the bifactor models as well because we did not want to confound the results by using different identification methods across models. Future studies of the WPPSI-IV might want to compare identification methods to see if that influences the results of the model fit.

Another possible limitation is the use of model fit indices to compare models. Barrett (2007) argued that the only acceptable statistical standard for model fit is the chi-square test and that "ad hoc "approximate fit" indices fail miserably" (p. 823). Concern about cutoff values for goodness-of-fit indices has also been expressed by other measurement experts (e.g., Marsh et al., 2005). More specifically, Murray and Johnson (2013) questioned the use of model fit indices in distinguishing between second-order and bifactor models and surmised that there might be a statistical bias favoring bifactor models because they are methodologically better able to account for unmodeled complexity (e.g., small cross-loadings) in the data. Given this potential bias, Murray and Johnson (2013) suggested that model selection should be based on "the specific aims of the studies and not on model fit" (p. 421). If the study aims to measure g then the two models do equally well. However, if the study aims to estimate domain-specific abilities, then the higher-order model produces domain-specific scores that reflect an amalgam of both *g* and domain-specific variance; consequently, "bifactor model factor scores should be preferred" (Murray & Johnson, 2013, p. 420) because they reflect the influence of domain-specific factors independent of *g*.

A final potential limitation is that this study relied on data from the standardization sample and it is unknown if the same structure will emerge among clinical samples, especially children with learning disorders and diverse language skills. It is also not known if the various WPPSI-IV index scores possess incremental predictive validity. One strength of a bifactor model is the ability to directly test whether the domain-specific factors predict external variables over and above a general factor. This cannot be accomplished with a typical higher-order model because the first-order factors carry variance from both first- and second-order factors (Brown, 2013; Chen et al., 2012). It was not possible to test incremental predictive validity in the current study, but analyses of other Wechsler scales have typically found little incremental predictive validity for domain-specific factors beyond *g* or the FSIQ score (Canivez, 2013; Glutting, Watkins, Konold, & McDermott, 2006; Parkin & Beaujean, 2012). However, replication and generalization studies should be conducted with the WPPSI-IV among a variety of clinical and nonclinical populations to better delineate the relationships between the general and domain-specific first-order factors and theory-relevant criteria (e.g., academic achievement) to provide additional guidance for practitioners.

References

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

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42,* 815–824. doi:10.1016/j.paid.2006.09.018

Bergeron, R., & Floyd, R. G. (2013). Individual part score profiles of children with intellectual disability: A descriptive analysis across three intelligence tests. *School Psychology Review, 42,* 22–38.

Betts, J., Pickart, M., & Heistad, D. (2011). Investigating early literacy and numeracy: Exploring the utility of the bifactor model. *School Psychology Quarterly, 26,* 97–107. doi:10.1037/a0022987

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment, 25,* 136–145. doi:10.1037/a0029228

Brown, T. A. (2013). Latent variable measurement models. In T. D. Little (Ed.), *Oxford handbook of quantitative methods: Statistical analysis* (Vol. 2, pp. 257–280). New York, NY: Oxford University Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs.

Journal of Personality, 80, 796–846. doi:10.1111/j.1467-6494.2011.00749.x

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. doi:10.1177/0049124104268644

Canivez, G. L. (2013). Incremental criterion validity of WAIS-IV factor index scores: Relationships with WIAT-II and WIAT-III subtest and composite scores. *Psychological Assessment, 25,* 484–495. doi:10.1037/a0032092

Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 1–17). Mahwah, NJ: Erlbaum.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80,* 219–251. doi:10.1111/j.1467-6494.2011.00739.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189–225. doi:10.1207/s15327906mbr4102_5

Coalson, D., & Weiss, L. G. (2002). The evolution of Wechsler intelligence scales in historical perspective. *Assessment Focus Newsletter, 11*(1), 1–3.

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299. doi:10.1037/1040-3590.7.3.286

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35,* 169–182. doi:10.1016/j.intell.2006.07.002

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 (Eds.), *Culture and children's intelligence* (pp. 277–313). San Diego, CA: Academic Press. doi:10.1016/B978-012280055-9/50021-7

Gignac, G. E. (2006). The WAIS-III as a nested factors model. *Journal of Individual Differences, 27,* 73–86. doi:10.1027/1614-0001.27.2.73

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *The Journal of Special Education, 40,* 103–114. doi:10.1177/00224669060400020101

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment, 23,* 143–152. doi:10.1037/a0021230

Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2013). Further insights on the French WISC-IV factor structure through Bayesian structural equation modeling. *Psychological Assessment, 25,* 496–508. doi:10.1037/a0030676

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Grégoire, J., Georgas, J., Saklofske, D. H., van de Vijver, F., Wierzbicki, C., Weiss, L. G., & Zhu, J. (2008). Cultural issues in clinical use of the WISC-IV. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (2nd ed., pp. 517–544). San Diego, CA: Academic Press.

Gustafsson, J.-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186–242). New York, NY: Macmillan.

Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika, 2,* 41–54. doi:10.1007/BF02287965

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

Keith, T. Z. (2005). 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* (2nd ed., pp. 581–614). New York, NY: Guilford Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59–72. doi:10.1207/s15328007sem1301_3

Liu, J., & Lynn, R. (2011). Factor structure and sex differences on the Wechsler Preschool and Primary Scale of Intelligence in China, Japan and United States. *Personality and Individual Differences, 50,* 1222–1226. doi:10.1016/j.paid.2011.02.013

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34,* 593–610. doi:10.1007/s10519-004-5587-0

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence, 41,* 407–422. doi:10.1016/j.intell.2013.06.004

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale-Fourth Edition with a clinical sample. *Psychological Assessment, 25,* 618–630. doi:10.1037/a0032086

Nelson, L. D. (2000). Introduction to the special section on methods and implications of revising assessment instruments. *Psychological Assessment, 12,* 235–236. doi:10.1037/1040-3590.12.3.235

Niileksela, C. R., Reynolds, M. R., & Kaufman, A. S. (2013). An alternative Cattell-Horn-Carroll (CHC) factor structure of the WAIS-IV: Age invariance of an alternative model for ages 70–90. *Psychological Assessment, 25,* 391–404. doi:10.1037/a0031175

Parkin, J., & Beaujean, A. A. (2012). The effects of Wechsler Intelligence Scale for Children—Fourth Edition cognitive abilities on math achievement. *Journal of School Psychology, 50,* 113–128. doi:10.1016/j.jsp.2011.08.003

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. doi:10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140. doi:10.1080/00223891.2012.725437

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16,* 19–31. doi:10.1007/s11136-007-9183-7

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287–297. doi:IO.J037111040-3590.12.3.287

Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, V. L. Schwean, & C. R. Reynolds (Eds.), *The Oxford handbook of child psychological assessment* (pp. 48–83). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199796304.013.0003

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67. doi:10.1207/s15327906mbr2301_3

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). La Mesa, CA: Sattler Publisher.

Shrout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (Vol. 1, pp. 643–660). Washington, DC: American Psychological Association. doi:10.1037/13619-034

Sivo, S. A., Xitao, F., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education, 74,* 267–288. doi:10.3200/JEXE.74.3.267-288

Spearman, C. (1927). *The abilities of man*. New York, NY: Cambridge.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment, 24,* 101–113. doi:10.1037/a0024712

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association. doi:10.1037/10694-000

Varni, J. W., Beaujean, A. A., & Limbers, C. A. (in press). Factorial invariance of pediatric patient self-reported fatigue across age and gender: A multigroup confirmatory factor analysis approach utilizing the PedsQL™ Multidimensional Fatigue Scale. *Quality of Life Research*. doi:10.1007/s11136-013-0370-4

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children-Fourth Edition among a national sample of referred students. *Psychological Assessment, 22,* 782–787. doi:10.1037/a0020043

Watkins, M. W. (2013). *Omega* [computer program]. Phoenix, AZ: Ed. & Psych Associates.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence—third edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—fourth edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—fourth edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence—fourth edition technical manual and interpretive manual*. San Antonio, TX: Psychological Corporation.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29,* 377–392. doi:10.1177/0734282911406668

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64,* 113–128. doi:10.1007/BF02294531

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123–133. doi:10.1007/s11336-003-0974-7

Received April 3, 2013
Revision received August 8, 2013
Accepted August 11, 2013 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://notify.apa.org/ and you will be notified by e-mail when issues of interest to you become available!