Psychological Assessment

Further Insights on the French WISC–IV Factor Structure Through Bayesian Structural Equation Modeling

Philippe Golay, Isabelle Reverte, Jérôme Rossier, Nicolas Favez, and Thierry Lecerf

Online First Publication, November 12, 2012. doi: 10.1037/a0030676

Psychological Assessment

1040-3590/12/$12.00 DOI: 10.1037/a0030676

Further Insights on the French WISC–IV Factor Structure Through Bayesian Structural Equation Modeling

Philippe Golay
University of Geneva and Distance Learning University, Switzerland

Isabelle Reverte
University of Geneva

Jérôme Rossier
University of Lausanne

Nicolas Favez and Thierry Lecerf
University of Geneva and Distance Learning University, Switzerland

The interpretation of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV) is based on a 4-factor model, which is only partially compatible with the mainstream Cattell–Horn–Carroll (CHC) model of intelligence measurement. The structure of cognitive batteries is frequently analyzed via exploratory factor analysis and/or confirmatory factor analysis. With classical confirmatory factor analysis, almost all cross-loadings between latent variables and measures are fixed to zero in order to allow the model to be identified. However, inappropriate zero cross-loadings can contribute to poor model fit, distorted factors, and biased factor correlations; most important, they do not necessarily faithfully reflect theory. To deal with these methodological and theoretical limitations, we used a new statistical approach, Bayesian structural equation modeling (BSEM), among a sample of 249 French-speaking Swiss children (8–12 years). With BSEM, zero-fixed cross-loadings between latent variables and measures are replaced by approximate zeros, based on informative, small-variance priors. Results indicated that a direct hierarchical CHC-based model with 5 factors plus a general intelligence factor better represented the structure of the WISC–IV than did the 4-factor structure and the higher order models. Because a direct hierarchical CHC model was more adequate, it was concluded that the general factor should be considered as a breadth rather than a superordinate factor. Because it was possible for us to estimate the influence of each of the latent variables on the 15 subtest scores, BSEM allowed improvement of the understanding of the structure of intelligence tests and the clinical interpretation of the subtest scores.

Keywords: WISC–IV, Bayesian structural equation modeling, direct hierarchical model, CHC theory

The last decade has seen the emergence of the Cattell–Horn–Carroll (CHC) theory as the mainstream approach in the field of intelligence assessment. This model was influential in the development and interpretation of several cognitive batteries, such as the KABC–II (Kaufman & Kaufman, 2008) and the Woodcock–Johnson–III (Woodcock, McGrew, & Mather, 2001). The CHC taxonomy is actually the best validated model of human cognitive abilities (Ackerman & Lohman, 2006). The current CHC framework is composed of three distinct stratums with a general factor of intelligence at the top, about 16 broad abilities in the middle, and about 100 narrow abilities on the bottom (McGrew, 2009; Newton & McGrew, 2010).

Since the creation of the Wechsler–Bellevue scale in 1939, the interpretation of the Wechsler intelligence scales has moved from a two-factor to a four-factor structure. The current version of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; Wechsler, 2005) distinguishes four index scores: Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI; see Figure 1a). The present study addressed three goals: Our first objective was to determine the most adequate model for describing the French WISC–IV. The second goal was to ascertain the exact nature of the constructs measured by each subtest score by estimating the relationship between every latent variable and subtest score. The third and final goal was to test the competing theories of superordinate general intelligence versus breadth general intelligence.

Even if the four factors of the WISC–IV are more attuned to contemporary theory, they were not aligned with the consensual CHC model of intelligence measurement (Flanagan, Ortiz, &

Philippe Golay, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland, and Faculty of Psychology, Distance Learning University, Sierre, Switzerland; Isabelle Reverte, Faculty of Psychology and Educational Sciences, University of Geneva; Jérôme Rossier, Institute of Psychology, University of Lausanne, Lausanne, Switzerland; Nicolas Favez, Faculty of Psychology and Educational Sciences, University of Geneva, and Faculty of Psychology, Distance Learning University, Switzerland; Thierry Lecerf, Faculty of Psychology and Educational Sciences, University of Geneva, and Faculty of Psychology, Distance Learning University, Switzerland.

This research was supported by Swiss National Science Foundation Grant 100014-118248 to Thierry Lecerf, Nicolas Favez, and Jérôme Rossier.

Correspondence concerning this article should be addressed to Philippe Golay, FPSE-Psychology, University of Geneva, 40 Boulevard du Pont d'Arve, CH-1205 Geneva, Switzerland. E-mail: philippe.golay@unige.ch

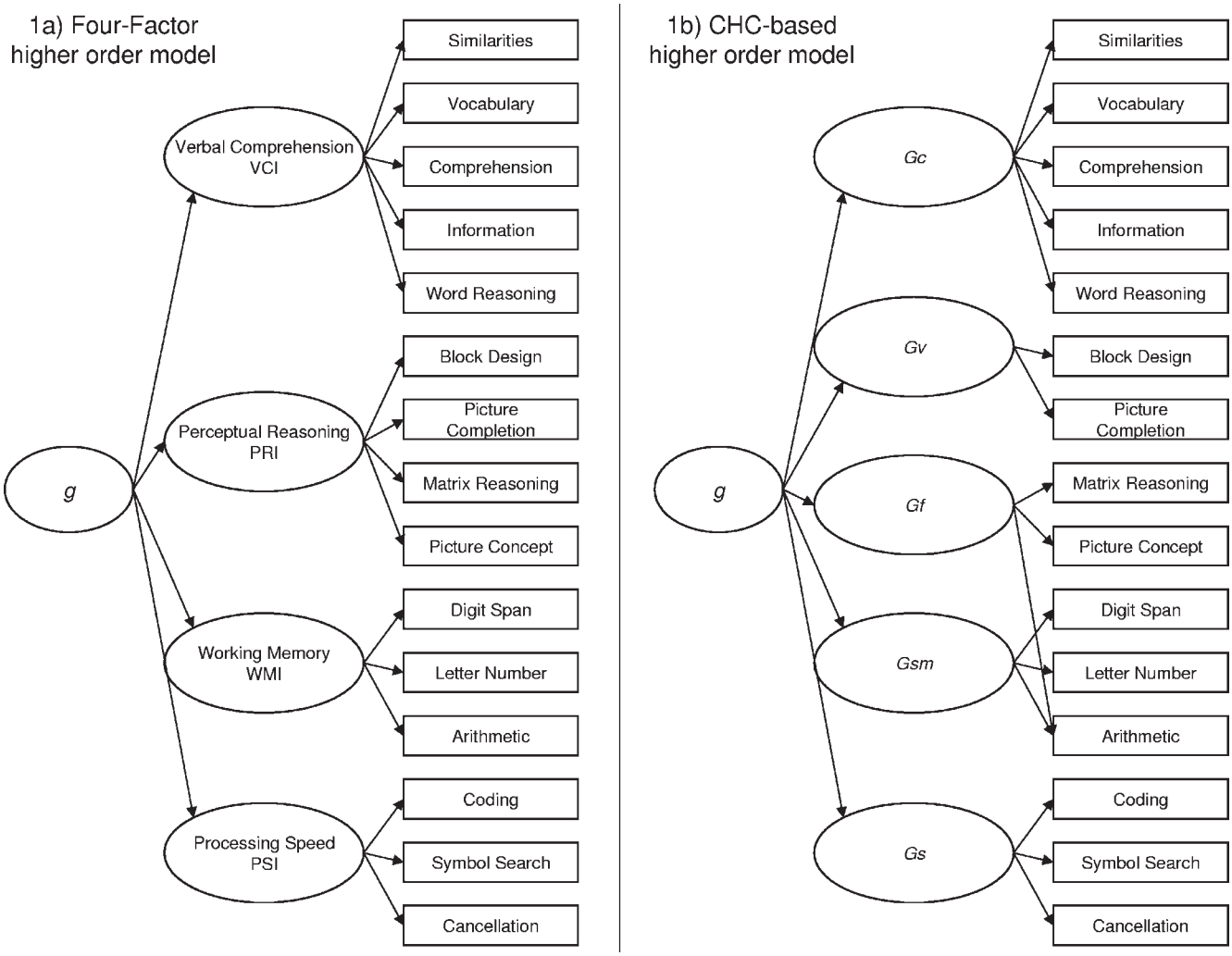

Figure 1. Four-factor higher order and CHC-based higher order models for the French WISC–IV. CHC = Cattell–Horn–Carroll; WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; *Gc* = comprehension-knowledge; *Gv* = visual processing; *Gf* = fluid reasoning; *Gsm* = short-term memory; *Gs* = processing speed.

Alfonso, 2007; Golay & Lecerf, 2011; Grégoire, 2009). However, confirmatory factor analytic studies conducted on the WISC–IV demonstrated that, in comparison to the standard four-factor structure, models based on the CHC framework better fitted the North American data (Keith, Fine, Taub, Reynolds, & Kranzler, 2006) and the French data (cf. Figure 1b; Lecerf, Rossier, Favez, Reverte, & Coleaux, 2010). It should be noted that confirmatory factor analyses conducted on the Wechsler Adult Intelligence Scale—Third Edition (WAIS–III) and the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV) also indicated that a CHC-based model was more adequate for both the North American data (Benson, Hulac, & Kranzler, 2010) and the French data (Golay & Lecerf, 2011). However, the four indexes of the WISC–IV could still be related to the CHC ability classification: VCI (Similarities, Vocabulary, Comprehension, Information, and Word Reasoning subtest scores) can be considered theoretically close or identical to the *Gc* factor (comprehension-knowledge) from the CHC framework, but PRI (Block Design, Picture Completion, Matrix Reasoning, and Picture Concepts subtest scores) is not represented per se in the CHC theory. According to the latter model, PRI should be split into both *Gf* (fluid reasoning: Matrix Reasoning plus Picture Concepts subtest scores) and *Gv* (visual processing: Block Design plus Picture Completion subtest scores) factors. Finally, WMI and PSI are related to respectively the *Gsm* (short-term memory) and the *Gs* (processing speed) CHC factors. Thus, the interpretation of some of the subtest scores and index scores may substantially differ depending on the reference theoretical model. Our first purpose in the present study was to compare the four-factor structure of the WISC–IV with a five-factor CHC-based model.

Following this first question regarding the factor structure of the WISC–IV, a second controversy remains on the nature of the constructs measured by each subtest score. Some of the subtests theoretically associated with one latent variable could in fact be associated with other latent variables (i.e., cross-loadings). This

issue brings into question the clinical utility and the interpretation of the subtest scores of the WISC–IV, because each subtest score could measure additional specific abilities. For instance, the Arithmetic subtest score has been shown to load on the short-term memory (*Gsm*) factor, on the fluid reasoning (*Gf*) factor, or on both factors (Keith et al., 2006). For the French version of the WAIS–III, the Arithmetic subtest score loaded on both the *Gf* factor and the *Gsm* factor (Golay & Lecerf, 2011), a finding that was consistent with results obtained with the French WISC–IV (Lecerf et al., 2010). Similarly, many other cross-loadings have been hypothesized: The Symbol Search subtest score may load not only on *Gs* but on *Gv*; the Word Reasoning score may measure *Gc* and *Gf*; the Block Design score may load on both *Gf* and *Gv,* and the Picture Completion score may load on both *Gv* and *Gc* (Keith et al., 2006). In line with these hypotheses, our second goal in this paper was to address the issue of the interpretation of the subtest scores in a straightforward manner, hence considering every cross-loading. The purpose was to determine whether secondary interpretation of some subtest scores is supported by the data. This is a critical issue because with confirmatory factor analysis, alternative models with different loadings and/or cross-loadings can show very close levels of fit to the data and ultimately all be considered plausible. Additionally, when trying to determine if some interpretation of the subtest scores is more correct than other alternatives, researchers may rely on multiple model modifications and comparisons, a process that may capitalize on chance. The methodology used in this paper, Bayesian structural equation modeling (BSEM), was chosen because it can deal with this methodological limitation by estimating all cross-loadings simultaneously.

Factor Analysis Techniques

Two basic types of factor analytic techniques have been used to evaluate the structure of intelligence and the factor structure of batteries like the Wechsler intelligence scales: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In brief, EFA is used when the relations between scores and latent variables are doubtful or unknown. EFA models are weakly specified, and the main question is to determine the number of factor that should be retained (even if several criteria could be used). With EFA, all loadings between measures and latent variables are freed and estimated simultaneously. Unlike in CFA, no factor loadings are fixed to zero. The necessary number of restrictions for model identification is achieved by rotating the factor-loading matrix and fixing factor variances. The main advantage of EFA is that "it lets the data speak for themselves" (Carroll, 1995, p. 436). Because all loadings are simultaneously estimated, EFA models are classically considered to be closer to the data than their CFA counterparts (Marsh et al., 2010).

In the majority of the recent studies examining the factor structure of the Wechsler intelligence scales or other intelligence tests, CFA instead of EFA was used. CFA is mainly used when the researchers have prior knowledge about the relationships between measures and latent variables. The main difference with EFA is that only some parameters of the model are estimated on the basis of theoretical and empirical considerations. CFA is most frequently used with a maximum-likelihood estimation procedure (ML-CFA), and it allows researchers to specify the measurement and the structural part of the model with greater flexibility. One another important advantage of CFA is the possibility of conducting statistical tests of different hypotheses and of assessing to what extent the model fit the data. These statistical goodness-of-fit indexes allow the researcher to compare alternative models and to select the hypothesized model that best represents the data. In some situations, the researcher can be tempted to test many modifications in order to achieve acceptable fit to the data. Alternative models can also demonstrate very close goodness of fit. However, because many models are tested, improvements are susceptible to capitalization on chance characteristics of the data: "A number of writers have cautioned that extensive specification searches can lead to unjustified overfitting of data, with loss of meaning for indices of statistical significance" (Carroll, 1995, p. 438). The selection of the best model on solid statistical ground may ultimately be illusory because different sets of hypotheses often appear to be equally plausible for a given data set (Carroll, 1995). Such a data-driven process casts doubt on whether the modification of the initial model could be generalized to other samples or to the population (MacCallum, Roznowski, & Necowitz, 1992; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). This practice might contribute to the lack of consensus between studies and is especially relevant to the question of the Wechsler scales when the influence of each of the latent variables on the subtest scores has to be estimated.

Most important, it should be mentioned that CFA imposes some restrictions in order to achieve model identification with standard estimation procedures like maximum-likelihood estimation. With typical ML-CFA, many cross-loadings between latent variables and measures are fixed to be exactly zero, which does not always faithfully reflect the researchers' hypotheses (Muthén & Asparouhov, 2012). Indeed, small but nonzero cross-loadings could be equally compatible with theory. In order to determine the exact nature of the constructs measured by each subtest score, it may be necessary to estimate many cross-loadings, one at a time. Furthermore, unnecessarily strict models and inappropriate zero cross-loadings can contribute to poor model fit, distorted factors, and biased factor correlations (Marsh et al., 2010). For instance, the last issue has raised a great debate in the domain of personality research: ML-CFA has been considered overly restrictive because the independent cluster model requires each indicator to load on one factor only (McCrae et al., 2008). In consequence, the model involves a large number of zero cross-loadings. Thus, residual correlations between the Big Five dimensions (supposed to be independent) have been considered to be method artifact: A modest relation between a specific item and a nontarget factor cannot be accounted for by the cross-loadings and is represented instead through the correlations between the factors. These factor correlations are then very likely to be overestimated (Marsh et al., 2010). It is fair to say that mediocre goodness of fit has not generally been an issue with ML-CFA conducted on the Wechsler scales. However, researchers are often left with the dilemma of whether to keep many meaningful alternatives untested or to risk overfitting their model to the data. BSEM, as is described below, offers a rather elegant solution to this limitation.

As mentioned, EFA and CFA are used to better understand the structure and the constructs measured by each subtest score of the WISC–IV. Nevertheless, in addition to the debate about the number of factors and the existence of some cross-loadings, there remains another debate about the structure of the Wechsler scales

and other intelligence tests: Developers claim that the French WISC–IV measures four cognitive dimensions beyond the general intelligence factor and that the four factor index scores are the primary level of score interpretation. However, several studies demonstrated that subtest scores reflected both second-order and first-order factors. It is therefore necessary to assess the relative importance of the general factor and index scores relative to subtest scores (Canivez & Watkins, 2010a, 2010b; Carroll, 1993; Gorsuch, 1983). Otherwise, there is a risk of confounding general and specialized abilities (Gustafsson & Balke, 1993). One solution is to use the Schmid and Leiman orthogonalization transformation (SLT) of oblique factors (Schmid & Leiman, 1957). This procedure, based on EFA, has been extensively used and advocated by Carroll in his seminal study of cognitive abilities (Carroll, 1993, 1995). The first-order factors are modeled orthogonally to each other and to the general factor. SLT allows one to derive a hierarchical factor model from higher order models and to decompose the variance of each subtest score into the general factor and the first-order factors. The SLT was applied to several intelligence tests—among them, the Stanford–Binet Intelligence Scales (SB-5), Reynolds Intellectual Assessment Scales (RIAS), Wide Range Intelligence Test (WRIT), Wechsler Abbreviated Scale of Intelligence (WASI), and WAIS–IV. Results indicated that most variance of the subtest scores was associated with the general factor and that the first-order factors accounted for small portions of the total and the common variance (Canivez, 2008; Canivez & Watkins, 2010a, 2010b; Golay & Lecerf, 2011; Watkins, 2006, 2010). It was concluded that Wechsler scales provided mainly a measure of general intelligence and that clinical interpretation should be primarily at that level.

When confirmatory factor analyses are used, as it was the case in the present study, direct hierarchical models are more adequate than higher order models for assessing the relative importance of the general factor and index scores relative to subtest scores (Gignac, 2008). With a direct hierarchical model, the *g* factor has no direct effects on the first-order factors, which are modeled orthogonally to each other (i.e., uncorrelated first-order factors), and all subtest scores directly load onto a general factor and also onto one first-order group factor. In contrast to higher order models, in which the relationship between the general factor and each subtest score is mediated only by the broad abilities, direct hierarchical models include a direct relationship between the general factor and each subtest score. Let us remember that Holzinger and Swineford (1937) first referred this model as the *bi-factor* model, whereas Gustafsson and Balke (1993) described it as a *nested factor model*. More recently, it was referred as the *direct hierarchical model* by Gignac (2008). The use of higher order or direct hierarchical models has substantive theoretical implications. Although the higher order models are more consistent with a superordinate conceptualization of *g*, direct hierarchical models involve a breadth conceptualization of *g* (Gignac, 2008; Golay & Lecerf, 2011; Watkins, 2010). Therefore, direct hierarchical models were tested in the present study to estimate the relative influence of the general factor and index scores on the WISC–IV subtest scores. Consequently, our third purpose in this study was to test the competing theories of higher order superordinate *g* versus breadth *g*. We assumed, in accordance with Gignac's results (2008), that our data would support a breadth conceptualization of *g* (direct hierarchical model) rather than a superordinate conceptualization of *g* (higher order model).

The Bayesian Structural Equation Modeling (BSEM) Approach

The method used here closely draws on the approach described by Muthén and Asparouhov (2012). This paragraph does not provide a detailed overviewed of BSEM estimation. Instead, we focus on some of the most important advantages of BSEM over standard maximum-likelihood CFA estimation. First of all, it is important to remember that ML-CFA and BSEM are not different statistical models: ML-CFA and BSEM represent two different estimation procedures used in the general context of confirmatory factor analysis. However, we will see that the Bayesian estimator used with BSEM allows one to test more complex structures.

CFA, whether estimated through ML-CFA or through BSEM, relies on the distinction between the *measurement model* and the *structural model*. The former specifies the relationships between the latent and the observed variables, and the latter specifies the relationships among the latent variables; both measurement and structural models reflect theory. In other words, knowledge based on previous studies is incorporated in the definition of the model. For instance, on the basis of the literature, a researcher could assume that the score of the subtest Symbol Search of the Wechsler Intelligence Scale measures the processing speed factor and should demonstrate little or no relation with the visualization factor.

However, it should be noted that ML and BSEM, the statistical estimation procedures used with CFA models, differ in several ways (cf. Table 1). First of all, although parameters are considered as constants with typical maximum-likelihood estimation, BSEM and Bayesian statistics see parameters as variables (Yuan & MacKinnon, 2009). The second and third main important differences between ML-CFA and BSEM are reflected in the parameter specification (cross-loadings and major loadings). With classical ML-CFA, the major loadings for parameters (i.e., processing speed to Symbol Search) are freely estimated, and the unexpected cross-loadings (i.e., visualization to Symbol Search) are fixed exactly to 0. Indeed, with ML-CFA, if all parameters were freed, the model would not be identified. BSEM differs from ML-CFA because the exact zero cross-loadings are replaced with "approximate zeros" based on informative prior distributions. This is particularly relevant in the present study because many cross-loadings could be suspected of being slightly greater than zero. Thus, in the first step of Bayesian estimation, the prior distributions for the parameters are based on previous studies (current knowledge) and reflect expectation as to the likely value of the parameters; nevertheless, there is still more or less uncertainty about the parameter values. Indeed, the prior is represented here by a normal distribution with a mean of zero, an infinite variance for the major loadings, and a small variance for cross-loadings:

> A non-informative prior, also called a diffuse prior, has a large variance. A large variance reflects large uncertainty in the parameter value. With a large prior variance the likelihood contributes relatively more information to the formation of the posterior and the estimate is closer to a Maximum-Likelihood estimate. (Muthén & Asparouhov, 2012, p. 315)

Table 1
Summary of ML-CFA and BSEM Differences

Difference	ML-CFA	BSEM
Parameters viewed as	Constants	Variables
Cross-loadings	Exact zeros	Informative priors (zero mean and small variance)
Major loadings	Freely estimated	Diffuse noninformative priors (zero mean and infinite variance)
Model modification	Improvement with modification indices one parameter at a time	All parameters freed and estimated simultaneously
Parameter estimates	Assumed to be normally distributed	Based on percentiles of the posterior distribution, does not assume a normal distribution

Note. ML-CFA = maximum-likelihood confirmatory factor analysis; BSEM = Bayesian structural equation modeling.

In sum, in order to estimate the model with BSEM, one can replace zero-fixed cross-loadings by normal prior distributions with a mean of zero and a small variance, which are approximate zeros. It should be noted that with classical ML-CFA, an exact zero cross-loading could be considered a very strong informative prior (with a mean and a variance of zero). Therefore, ML-CFA is more restrictive because there is absolutely no uncertainty on these parameters. The models tested with ML-CFA are theoretically driven, but the zero-fixed cross-loadings are considered here as unnecessarily strict operationalization of the researcher's hypotheses. Because BSEM allows some degree of uncertainty in the parameters, this estimation technique is less restrictive than ML-CFA. However, BSEM still encompasses strong theory because the basic framework of the analysis is still confirmatory in nature. Prior knowledge is taken into account because the researcher must specify the model and must also provide prior distributions for the model parameters, including cross loadings.

After model specification, Bayesian estimation combines prior distributions for parameters with new collected data and forms posterior distributions for parameters through the Bayes theorem (for a detailed summary, see Yuan & MacKinnon, 2009). In other words, the posterior distribution is the updated representation of the researcher's belief, after incorporation of the experimental data, and provides Bayesian estimates. Information about model fit is obtained by using posterior predictive checking (PPC; Gelman, Meng, & Stern, 1996). A low PPC is considered to reflect a poor fit; a PPC value around 0.5 is considered to reflect a very good fit (see below).

The fourth important difference between ML-CFA and BSEM estimation concerns the modification of the initial model and the use of the modification indices derived from ML-CFA (see Table 1). Modification indices inform about model improvement when only one parameter is freed at a time. This could be an issue when the model fit is bad and many model modifications must be tested in order to achieve acceptable fit. This also could be problematic when several meaningful cross-loadings are suspected of being other than zero and many models should be tested, one at a time. In contrast, with Bayesian estimation of CFA, all parameters can be freed and estimated simultaneously, and therefore no modifications have to be tested. Thus, BSEM allow a simpler procedure for testing alternative cross-loadings, because all parameters are estimated at the same time.

A fifth and last important difference between ML-CFA and BSEM concerns the parameter estimates (see Table 1). ML-CFA estimation is based on asymptotic large-sample normality of the maximum-likelihood estimates. Therefore, it is not appropriate to use asymptotic theory with ML-CFA when the sample size is small, because the sampling distribution of parameter estimates is unknown and often cannot be estimated efficiently by applying formulas based on asymptotic theory (Scheines, Hoijtink, & Boomsma, 1999). In contrast, BSEM does not rely on large sample theory (Lee & Song, 2004). The parameter estimates and the credibility intervals are based on the percentile of the posterior distribution. Parameters are considered to have substantive backing when the 95% credibility interval of the parameter does not cover zero (Muthén & Asparouhov, 2012). Therefore, it is not necessary to rely on normal approximations of the posterior (Scheines et al., 1999). Thus, BSEM has the advantage of being able to accommodate heavily skewed distributions of parameter estimates. Finally, some authors have shown that Bayesian estimation does better than ML when sample size is small (Lee & Song, 2004).

Method

Participants

Participants included 249 French-speaking Swiss children (124 male, mean age = 9.69 years, SD = 1.18, and 125 female, mean age = 9.78 years, SD = 1.20). The children were recruited from several schools of the Canton of Geneva (Switzerland), and all children were in the school grade appropriate to their chronological age. The sample was stratified according to sex and parents' education level, closely approximating the Geneva Census.

Instrument and Procedure

The French version of the WISC–IV (Wechsler, 2005) is an individually administered test of intelligence that includes 10 core and five supplemental subtests. The version we used in the present study with French-speaking Swiss children was standardized on a French nationally representative sample of children from 6 to 16.11 years of age (N = 1,103). The French WISC–IV provides four-factor-based indexes: Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI), and Processing Speed (PSI). The Full Scale IQ is based on the addition of the 10 core subtest scores. The standard scores (M = 10, SD = 3) of the 15 subtests were used to conduct analyses, and there were no missing data. Subtest standard scores were standardized so that the scale of the priors corresponds to standardized loadings.

Models and Analyses

First, in order to analyze the structure of the WISC–IV, we used BSEM to compare the current four-factor higher order structure of the WISC–IV with a five-factor CHC-based model. Second, BSEM was used to improve the understanding of the constructs measured by the French WISC–IV subtest scores by allowing all cross-loadings to be estimated at the same time. Thus, the influence of each latent variable on the 15 subtest scores was estimated, which is not possible with classical ML-CFA estimation. The third and last purpose was to compare direct hierarchical models with more classical higher order models, because the former have been found to show a better fit to the data. The comparison between higher order and direct hierarchical models allowed us to test the competing theories of higher order superordinate *g* versus breadth *g*.

In sum, we compared eight models: The first four models were based on the current four-factor structure. Models 1 and 2 were higher order models with cross-loadings either fixed to zero or freely estimated (see Figure 1a). The objective was to estimate the gains in term of model fit that could be achieved. Models 3 and 4 were direct hierarchical counterparts of Models 1 and 2 (see Figure 2a). Then, CHC-based models with five factors were tested: Models 5 and 6 were higher order models with cross-loadings either fixed to zero or freely estimated (see Figure 1b). Models 7 and 8 were their direct hierarchical alternatives (see Figure 2b).

As described, BSEM follows the same rationale as typical ML-CFA but includes two additional steps for the definition of the model to be estimated. For the WISC–IV, the first step was to determine the measurement and the structural parts of the models (i.e., define which subtest scores should be associated with which factors). For instance, Coding, Symbol Search, and Cancellation were set to load on the processing speed factor (PSI). With typical ML-CFA, the cross-loadings of all other subtest scores on the PSI factor should have been fixed to zero. With BSEM, cross-loadings

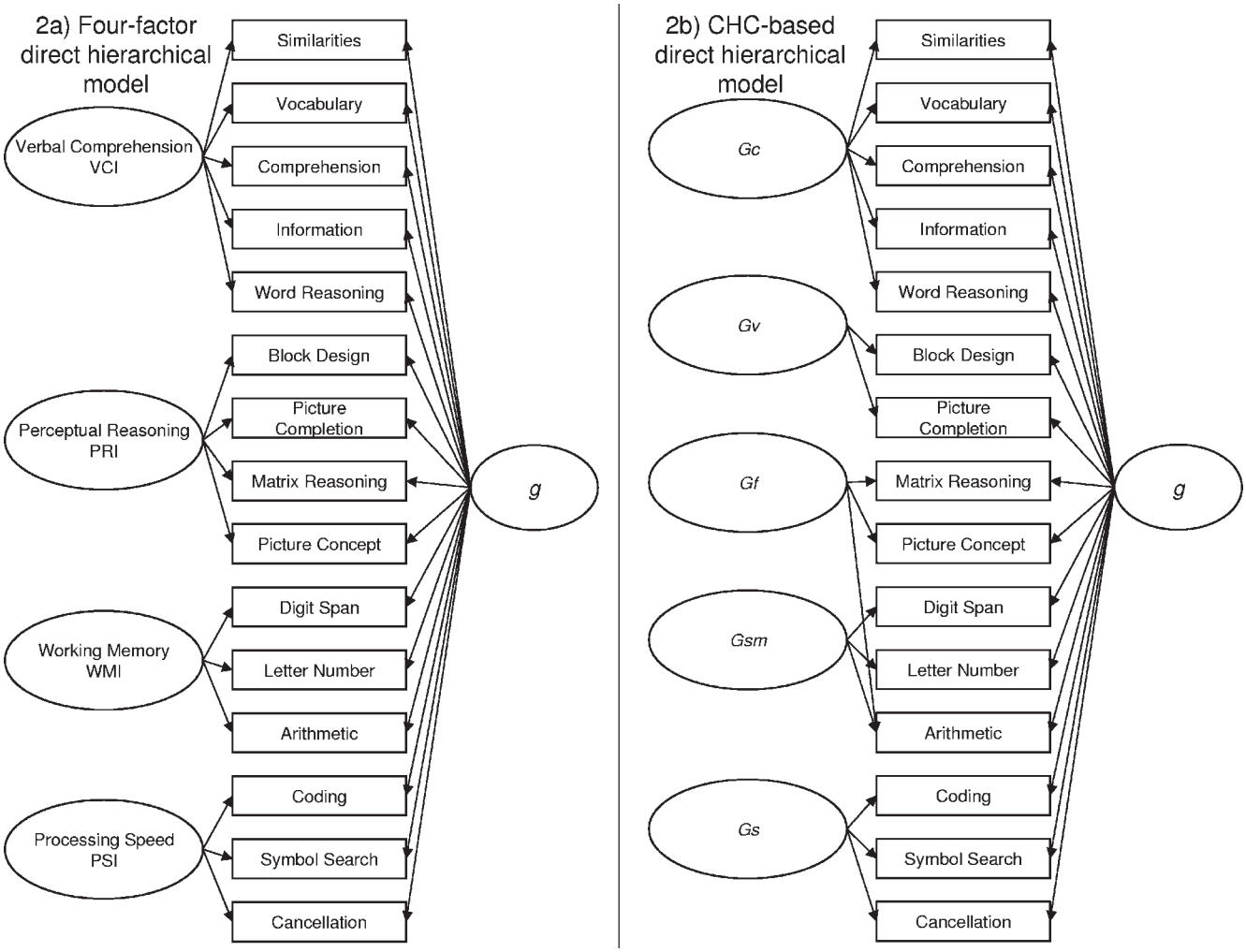

Figure 2. Four-factor direct hierarchical and CHC-based direct hierarchical models for the French WISC–IV. CHC = Cattell–Horn–Carroll; WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; *g* = general factor; *Gc* = comprehension-knowledge; *Gv* = visual processing; *Gf* = fluid reasoning; *Gsm* = short-term memory; *Gs* = processing speed.

can be specified as prior distributions of a mean of zero and a small variance. In that case, the prior is informative in the sense that all cross-loading values are small but are not necessarily zero. Each subtest score was therefore allowed to load on all factors. Using informative, small-variance priors for all cross-loadings would give information on the model, which would not be identified otherwise. The choice of the prior variance reflects initial beliefs and theoretical knowledge. A very small variance would not allow some cross-loadings to differ sufficiently from zero. On the other hand, a too large variance would not give enough information, so that the model would get closer to being nonidentified (Muthén & Asparouhov, 2012). In such cases, the Markov chain Monte Carlo (MCMC) estimation algorithm, which is the computational algorithm used with BSEM, will fail to reach convergence. In order to identify potential significant cross-loadings that remain rather inconclusive from the literature, we decided to choose the prior variance of 0.04 for the higher order models. The value of 0.04 results in 95% credibility interval of ±0.39; thus, this prior variance value allows small to moderate loadings. With direct hierarchical models, the standardized coefficients from *g* to subtest scores are typically higher than the standardized coefficients from the broad first-order factors to subtest scores (Gignac, 2006; Golay & Lecerf, 2011); therefore, we chose a prior variance of 0.01, which results in a 95% credibility interval of ±0.20.

Note that we also followed the recommendation of Muthén and Asparouhov (2012) concerning studying the sensitivity of the results: Smaller and larger prior variances were tested for each model and showed similar pattern of results, although slightly differing in regard to the model fit (i.e., lower posterior predictive *p* values; PPP). With larger prior variance values (0.05 for the higher order models and 0.02 for the direct hierarchical models), the MCMC process failed to reach convergence (models were not identified).

Posterior Distribution, MCMC, and Convergence

All statistical analyses were performed with the Mplus statistical package (version 6.11, Muthén & Muthén, 2010). The posterior distribution of Bayesian estimation was achieved through the MCMC algorithm with the Gibbs sampler method: "The idea behind MCMC is that the conditional distribution of one set of parameters given other sets can be used to make random draws of parameters values, ultimately resulting in an approximation of the joint distribution of all parameters" (Muthén & Asparouhov, 2012, p. 334). We did not use thinning, which consist of estimating the posterior distribution on the basis of every *k*th iteration rather than every iteration in order deal with potential autocorrelation in the chain. Thinning has often been considered as unnecessary and inefficient, given that the number of iterations is sufficiently large (Link & Eaton, 2011; MacEachern & Berliner, 1994; Muthén & Asparouhov, 2012).

Three MCMC chains with 50,000 iterations and with different starting values and different random seeds were used, a procedure that makes it possible to monitor convergence. As recommended by Muthén and Asparouhov (2012), the convergence was assessed with the Gelman–Rubin convergence diagnostic, which takes into account the potential scale reduction factor (PSR; Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Rubin, 1992). When PSR is between 1 and 1.1, convergence is considered to have been achieved, because the value indicates that the between-chain variation is small relative to the within-chain variation. For each model, we verified convergence by checking the PSR values (<1.1). Because it was important that the number of iterations was sufficiently large, we checked that the PSR values had converged in the desired range before the first half (25,000) of the iterations. This was easily achieved for the higher order models, but it was not always the case for the direct hierarchical models (PSR below 1.1 between 25,000 and 35,000 iterations). Thus, the direct hierarchical models were re-estimated with 70,000 iterations, which yielded essentially the same results. Finally, the first half of the chains was discarded as a burn-in phase, and the second half was used to estimate the posterior distribution (Muthén & Muthén, 2010).

Comparison of Model Fit

Model fit was assessed with posterior predictive checking (Gelman et al., 1996). The posterior predictive *p* value (PPP) of model fit is computed and can be used to test the structural model for misspecification. A small positive value (e.g., 0.005) indicates poor fit, and a PPP value around 0.5 indicates excellent fit (Muthén & Asparouhov, 2012). In contrast with standard ML-CFA goodness-of-fit indices, such as the root-mean-square error of approximation (RMSEA), there is no clear-cut PPP value that may indicate whether or not model fit is acceptable. Therefore, PPP should be interpreted like a structural equation modeling fit index: A bigger PPP indicates a better model. We additionally used the deviance information criterion (DIC; Gelman et al., 2004; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). The DIC is a Bayesian generalization of the AIC that balances the largeness of the likelihood and adds a penalty for model complexity (number of parameters). The number of parameters used to penalize for model complexity with the DIC is the effective number of parameters, referred as pD. Models with smaller values of DIC should be preferred.

Results

Four-Factor Models

The higher order four-factor models with and without cross-loadings were tested first. As shown in Table 2, Model 1 was clearly rejected (PPP = 0.005). When cross-loadings were allowed to differ from zero (Model 2), the fit dramatically increased (PPP = 0.456). The value of the DIC in Model 2 (9,616.227) was also lower than in Model 1 (9,645.043), indicating that approximate zero cross-loadings provided a better explanation of the structure of the WISC–IV than did the current factorial structure with cross-loadings fixed to 0.

Model 3 was a direct hierarchical variant of Model 1, with four first-order factors and one general factor. As shown in Table 2, Model 3 was also rejected because of a very low PPP value (0.072). Model 4 was a direct hierarchical model with freely estimated cross-loadings. As indicated by a PPP value of 0.569 and a DIC of 9,491.086, Model 4 provided a better description of the French WISC–IV subtest scores. It should be noted that the loadings of the four factors on *g* were substantially reduced when the cross-loadings were freely estimated. This finding suggests that

Table 2
Comparisons of Model Fit for the French WISC–IV Structure

Model	No. free parameters	PPP	Difference between observed & replicated χ^2 95% CI		DIC	pD
			Lower 2.5%	Upper 2.5%		
1. Standard WISC–IV, higher order	49	0.005	13.262	90.287	9,645.043	47.335
2. Standard WISC–IV, higher order + cross-loadings (priors variance = 0.04)	94	0.456	−39.380	43.376	9,616.227	68.984
3. Standard WISC–IV, direct hierarchical	60	0.072	−10.437	68.962	9,594.894	19.765
4. Standard WISC–IV, direct hierarchical + cross-loadings (priors variance = 0.01)	105	0.569	−47.814	37.672	9,491.086	−49.830
5. CHC model, higher order	51	0.028	−0.993	75.815	9,632.226	49.016
6. CHC model, higher order + cross-loadings (priors variance = 0.04)	110	0.529	−43.061	38.097	9,601.477	58.621
7. CHC model, direct hierarchical	61	0.046	−5.808	73.848	9,615.542	35.416
8. CHC model, direct hierarchical + cross-loadings (priors variance = 0.01)	120	0.620	−51.153	35.994	9,493.646	−44.408

Note. WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; CHC = Cattell–Horn–Carroll; PPP = posterior predictive *p* value; CI = credibility interval; DIC = deviance information criterion; pD = estimated number of parameters.

fixing small cross-loadings between specific items and nontarget factors to zero tends to inflate the factor correlations (Marsh et al., 2010). Although not reported here, for Model 2, the 95% credibility interval for the loading of PSI on *g* failed to exclude zero. This finding indicates that, when adopting a four-factor framework, correlations between PSI and the three other broad factors are probably overestimated.

CHC-Based Models

Next, we tested a CHC-based model with cross-loadings fixed to zero (Model 5) and another CHC-based model in which cross-loadings were freely estimated (small variance priors; Model 6). The comparison of Model 1 and Model 5 supports the hypothesis that the CHC framework provides a better description of the WISC–IV subtest scores than the classical four-factor model. As shown in Table 3, only one cross-loading was significant: Arithmetic loaded on both the *Gf* factor and the *Gsm* factor. However, this model showed poor fit (PPP = 0.028, DIC = 9,632.226).

Free estimation of the cross-loadings (Model 6) greatly improved the fit of the model (PPP = 0.529, DIC = 9,601.477; see Table 2). As shown in Table 4, all hypothesized major loadings had substantive backing, except the loading from *Gf* to Arithmetic, which was very low and did not exclude zero (0.115, 95% CI −0.319 to −0.519). In line with our previous classical ML-CFA analyses, this finding suggests that the Arithmetic score is not a measure of *Gf* in the French WISC–IV (Lecerf et al., 2010). It should be emphasized that the typical CFA approach to model specification (as represented in Model 5) suggests adopting a more complicated model (with the Arithmetic score loading on both *Gf* and *Gsm*), whereas the BSEM approach (Model 6) suggests a simpler model, in which Arithmetic loads only on *Gsm*. No other cross-loading were considered to be substantive, because the 95% CI always included zero. Although these nonzero cross-loadings do not possess much clinical significance (because their magnitude is small), they are nevertheless statistically important for a correct and nonbiased estimation of the model parameters. The small magnitudes of the cross-loadings suggest that besides the major loadings, no secondary interpretation of the subtest scores is backed up by the data.

The saturations of the first-order factors in Model 6 were reduced in comparison to those found in Model 5, suggesting that inappropriate fixed zero cross-loadings increased first-order factor correlations and thus loadings on the *g* factor. Indeed, the loading of *Gf* on *g* was estimated at .956 in Model 5 (see Table 3) and .880 in Model 6 (see Table 4). Although it is not reported here, we found with maximum-likelihood estimation that the loading of *Gf* on *g* was unitary. This finding demonstrates again that inappropriate fixed zero cross-loadings likely increased first-order factor correlations and thus loadings on the *g* factor; more precisely, the second order loadings are often overestimated and are sometimes even found to be unitary with classical ML-CFA analyses.

Finally, we estimated direct hierarchical CHC models. Once again, the model with cross-loadings fixed to zero (Model 7) was rejected (PPP = 0.046). However, the direct hierarchical five-factor model with freely estimated cross-loadings (Model 8) showed the best overall fit to the data (PPP = 0.620, DIC = 9,493.646). As shown in Table 5, the direct loadings of the subtest scores on *g* were all substantively backed up. However, it is important to note that only the *Gsm* and *Gs* factors were really supported. Indeed, when the variance of the general factor was extracted as a breadth factor, *g* explained more variance of the subtest scores than did *Gc*, *Gv,* and *Gf*. Most important, no 95% credibility interval of the subtest loadings on *Gc*, *Gv,* and *Gf* excluded zero. These results suggest that there is very weak support for calculation of separate *Gc*, *Gv,* and *Gf* factor scores. In contrast, *Gs* and *Gsm* showed a distinct contribution in addition to the part of variance explained by the general factor.

Discussion

Analyses reported in the technical manual of the WISC–IV support a four-factor structure. However, this factorial structure

Table 3
BSEM Analysis With No Cross-Loadings, CHC Model (Model 5)

First order loadings estimates (median)	*Gc* 95% CI	*Gv* 95% CI	*Gf* 95% CI	*Gsm* 95% CI	*Gs* 95% CI
Similarities	**0.767 0.698 0.823**	0	0	0	0
Vocabulary	**0.812 0.751 0.861**	0	0	0	0
Comprehension	**0.665 0.578 0.737**	0	0	0	0
Information	**0.803 0.741 0.853**	0	0	0	0
Word Reasoning	**0.708 0.628 0.773**	0	0	0	0
Block Design	0	**0.641 0.494 0.787**	0	0	0
Picture Completion	0	**0.556 0.412 0.690**	0	0	0
Matrix Reasoning	0	0	**0.601 0.480 0.712**	0	0
Picture Concepts	0	0	**0.468 0.345 0.586**	0	0
Digit Span	0	0	0	**0.618 0.492 0.730**	0
Letter–Number Sequencing	0	0	0	**0.761 0.635 0.890**	0
Arithmetic	0	0	**0.354 0.145 0.531**	**0.337 0.147 0.539**	0
Coding	0	0	0	0	**0.583 0.431 0.711**
Symbol Search	0	0	0	0	**0.769 0.623 0.935**
Cancellation	0	0	0	0	**0.450 0.304 0.582**

Second order loadings estimates (median)	General factor *g* 95% CI
Gc	**0.780 0.670 0.877**
Gv	**0.751 0.577 0.902**
Gf	**0.956 0.827 0.998**
Gsm	**0.592 0.430 0.735**
Gs	**0.427 0.261 0.576**

Note. Loadings in bold were freely estimated. Other loadings were fixed to 0. BSEM = Bayesian structural equation modeling; CHC = Cattell–Horn–Carroll; *Gc* = comprehension-knowledge; *Gv* = visual processing; *Gf* = fluid reasoning; *Gsm* = short-term memory; *Gs* = processing speed; CI = credibility interval.

is not explicitly aligned with the largely consensual Cattell–Horn–Carroll (CHC) theory of cognitive ability. Nevertheless, it has been shown that the WISC–IV could be better described with a CHC-based model, including five factors (Flanagan & Kaufman, 2009; Keith et al., 2006). Some questions also remain about the factorial structure of the WISC–IV and about the constructs measured by each subtest score. The majority of the studies addressing these questions used confirmatory factor analysis with a maximum-likelihood estimation procedure. However, this approach has several methodological and theoretical limits.

To address some of these limitations, we examined the French WISC–IV structure using Bayesian structural equation modeling (BSEM). The present study deals with three main questions: (a) Does a five-factor CHC-based model better fit the WISC–IV data than does the current four-factor structure? (b) What constructs are measured by each subtest score, and should cross-loadings be added to the model? Remember that with BSEM, the estimation procedure allows one to test for more complex model structures because all parameters are free and are estimated simultaneously (major loadings and cross-loadings). This may result in a final model that is less sample specific and hence more suited to other samples. (c) Does a direct hierarchical model better fit the data than does a high-order model, supporting a breadth conceptualization of *g* rather than a superordinate conceptualization of *g*?

Table 4
BSEM Analysis With Small-Variance Priors for Cross-Loadings, CHC Model (Model 6)

First order loadings estimates (median)	*Gc* 95% CI	*Gv* 95% CI	*Gf* 95% CI	*Gsm* 95% CI	*Gs* 95% CI
Similarities	**0.631**	0.130	0.045	0.018	0.000
	0.461 0.761	−0.037 0.318	−0.071 0.248	−0.135 0.178	−0.138 0.139
Vocabulary	**0.806**	−0.061	0.010	0.059	0.023
	0.656 0.976	−0.234 0.112	−0.130 0.174	−0.097 0.221	−0.116 0.162
Comprehension	**0.722**	−0.117	−0.007	0.049	−0.007
	0.569 0.901	−0.298 0.052	−0.166 0.136	−0.106 0.215	−0.151 0.134
Information	**0.743**	0.136	−0.021	0.004	−0.003
	0.586 0.917	−0.031 0.324	−0.208 0.109	−0.149 0.163	−0.141 0.137
Word Reasoning	**0.661**	0.045	0.031	0.011	−0.036
	0.503 0.817	−0.126 0.216	−0.091 0.228	−0.143 0.163	−0.178 0.099
Block Design	−0.069	**0.665**	0.004	0.037	0.126
	−0.260 0.127	**0.430 0.929**	−0.181 0.162	−0.164 0.221	−0.041 0.291
Picture Completion	0.107	**0.530**	0.013	−0.098	0.002
	−0.088 0.299	**0.287 0.810**	−0.128 0.195	−0.290 0.069	−0.157 0.153
Matrix Reasoning	0.039	0.139	**0.500**	0.023	−0.028
	−0.183 0.246	−0.094 0.356	**0.048 0.920**	−0.182 0.212	−0.193 0.124
Picture Concepts	0.071	−0.037	**0.385**	0.096	0.041
	−0.137 0.260	−0.256 0.153	**0.083 0.756**	−0.090 0.276	−0.109 0.188
Digit Span	−0.025	−0.083	0.022	**0.642**	0.021
	−0.209 0.151	−0.283 0.108	−0.114 0.215	**0.435 0.873**	−0.137 0.181
Letter–Number Sequencing	0.023	0.032	0.001	**0.749**	−0.111
	−0.173 0.212	−0.171 0.244	−0.155 0.173	**0.528 1.011**	−0.278 0.045
Arithmetic	0.128	−0.004	**0.115**	**0.400**	0.106
	−0.064 0.308	−0.198 0.191	**−0.319 0.519**	**0.173 0.690**	−0.037 0.254
Coding	−0.078	−0.007	0.001	0.040	**0.630**
	−0.251 0.095	−0.202 0.199	−0.154 0.161	−0.133 0.231	**0.456 0.809**
Symbol Search	0.002	0.106	−0.007	0.035	**0.657**
	−0.174 0.178	−0.089 0.313	−0.169 0.148	−0.141 0.215	**0.483 0.861**
Cancellation	0.054	0.017	0.015	−0.131	**0.494**
	−0.109 0.217	−0.163 0.213	−0.124 0.184	−0.304 0.029	**0.326 0.668**

Second order loadings estimates (median)	General factor *g* *95% CI*
Gc	**0.725**
	0.352 0.915
Gv	**0.694**
	0.248 0.937
Gf	**0.880**
	0.285 0.996
Gsm	**0.623**
	0.236 0.905
Gs	**0.386**
	−0.072 0.706

Note. Loadings in bold were freely estimated. Other loadings were estimated with small (0.04) variance priors. BSEM = Bayesian structural equation modeling; CHC = Cattell–Horn–Carroll; *Gc* = comprehension-knowledge; *Gv* = visual processing; *Gf* = fluid reasoning; *Gsm* = short-term memory; *Gs* = processing speed; CI = credibility interval.

First, concerning the debate about the structure of the WISC–IV, the results of the present investigation clearly indicate that the French WISC–IV subtest scores could be better described with a second-order general intelligence factor and five first-order factors defined according to the CHC framework: fluid reasoning (*Gf*), comprehension-knowledge (*Gc*), visual processing (*Gv*), processing speed (*Gs*), and short-term memory (*Gsm*). This finding is congruent with previous ML-CFA analyses and suggests that the CHC model would allow clinicians to make more adequate interpretations than the standard four-factor structure (Keith et al., 2006; Lecerf et al., 2010). Therefore, to gain clinical validity of tests scores interpretation, we recommend interpreting the results of the French WISC–IV subtest scores according to the CHC classification, with the norms that we have developed for the French WISC–IV (Lecerf et al., 2012). The superiority of the direct hierarchical and the higher order CHC-based models over the current four-factor alternatives also reinforces the idea that the PRI score should be split into two subcomponents (*Gf* and *Gv*).

Second, following this first debate about the factorial structure of the French WISC–IV, results from this investigation provide some very important information about the interpretation and constructs measured by the score of each subtest. In brief, the main results were as follows: All but one of the major loadings depicted on Figure 1b were supported by the data, as the Arithmetic subtest score loaded on the *Gsm* factor but not on the *Gf* factor. Remember that Keith et al. (2006) showed that the subtest measured *Gf* (and

Table 5
BSEM Analysis With Small-Variance Priors for Cross-Loadings, CHC Direct Hierarchical Model (Model 8)

Loadings estimates (median)	*G* 95% CI	*Gc* 95% CI	*Gv* 95% CI	*Gf* 95% CI	*Gsm* 95% CI	*Gs* 95% CI
Similarities	**0.665**	**0.314**	−0.012	−0.002	−0.012	−0.013
	0.506 0.791	**−0.523 0.557**	−0.166 0.152	−0.165 0.161	−0.138 0.111	−0.131 0.101
Vocabulary	**0.636**	**0.448**	0.015	0.000	0.035	0.017
	0.457 0.809	**−0.647 0.690**	−0.161 0.179	−0.157 0.156	−0.093 0.160	−0.101 0.134
Comprehension	**0.494**	**0.405**	0.033	0.000	0.038	−0.011
	0.300 0.676	**−0.611 0.638**	−0.192 0.209	−0.153 0.153	−0.093 0.166	−0.130 0.111
Information	**0.667**	**0.380**	−0.019	−0.003	−0.017	−0.017
	0.491 0.816	**−0.603 0.640**	−0.194 0.176	−0.215 0.212	−0.145 0.111	−0.137 0.100
Word Reasoning	**0.603**	**0.305**	0.003	−0.001	−0.018	−0.048
	0.434 0.742	**−0.515 0.549**	−0.144 0.147	−0.166 0.166	−0.146 0.105	−0.166 0.069
Block Design	**0.491**	−0.033	**−0.308**	−0.001	0.019	0.104
	0.306 0.662	−0.216 0.180	**−0.827 0.804**	−0.189 0.187	−0.138 0.171	−0.040 0.241
Picture Completion	**0.446**	0.013	**−0.176**	−0.001	−0.101	−0.014
	0.279 0.597	−0.159 0.178	**−0.574 0.567**	−0.170 0.168	−0.240 0.037	−0.148 0.118
Matrix Reasoning	**0.540**	0.005	−0.034	**−0.004**	0.030	−0.020
	0.398 0.667	−0.148 0.156	−0.227 0.207	**−0.396 0.396**	−0.112 0.164	−0.151 0.106
Picture Concepts	**0.444**	0.013	0.013	**−0.004**	0.073	0.035
	0.282 0.595	−0.159 0.178	−0.176 0.194	**−0.618 0.607**	−0.078 0.211	−0.098 0.163
Digit Span	**0.342**	−0.004	0.016	−0.002	**0.566**	0.026
	0.166 0.514	−0.156 0.149	−0.175 0.191	−0.182 0.178	**0.342 0.779**	−0.108 0.158
Letter–Number	**0.458**	0.007	−0.008	0.000	**0.552**	−0.090
Sequencing	**0.295 0.612**	−0.145 0.159	−0.167 0.162	−0.169 0.170	**0.332 0.767**	−0.220 0.039
Arithmetic	**0.523**	0.027	0.005	**−0.001**	**0.294**	0.078
	0.373 0.667	−0.174 0.199	−0.150 0.162	**−0.421 0.407**	**0.085 0.463**	−0.051 0.205
Coding	**0.188**	−0.019	−0.003	−0.002	0.046	**0.632**
	0.006 0.364	−0.184 0.160	−0.161 0.161	−0.166 0.162	−0.099 0.188	**0.460 0.804**
Symbol Search	**0.360**	−0.009	−0.019	−0.001	0.012	**0.573**
	0.191 0.518	−0.161 0.144	−0.199 0.181	−0.187 0.182	−0.127 0.150	**0.403 0.736**
Cancellation	**0.178**	0.016	−0.007	0.000	−0.106	**0.451**
	0.005 0.342	−0.155 0.177	−0.163 0.156	−0.180 0.179	−0.240 0.029	**0.286 0.605**

Note. Loadings in bold were freely estimated. Other loadings were estimated with small (0.01) variance priors. BSEM = Bayesian structural equation modeling; CHC = Cattell–Horn–Carroll; *G* =general factor; *Gc* = comprehension-knowledge; *Gv* = visual processing; *Gf* = fluid reasoning; *Gsm* = short-term memory; *Gs* = processing speed; CI = credibility interval.

also *Gsm* and *Gc*). On the other hand, Flanagan and Kaufman (2009) assumed that Arithmetic required *Gf* and *Gsm,* particularly for younger children. Concerning the French version of the WISC–IV, Grégoire (2009) suggested that the Arithmetic score measured both *Gsm* and *Gc*. Our previous ML-CFA indicated that the French Arithmetic subtest score measured quantitative knowledge (*Gq*) and short-term memory (*Gsm*) and, most important, that it did not measure fluid reasoning (*Gf*). The results found with BSEM were relatively consistent with those results: The Arithmetic score loaded only on *Gsm* (and did not load on *Gf*). Therefore, the Arithmetic score should be considered as a measure of *Gsm*, along with Digit Span and Letter–Number Sequencing. Nevertheless, the magnitude of the loading of Arithmetic on *Gsm* was smaller than the loadings of Digit Span and Letter–Number Sequencing on this factor. It should be noted that we did not test the hypothesis that the Arithmetic score measures quantitative knowledge (*Gq*) in the present study, because Arithmetic was the only subtest score that appears to measure this ability in the French WISC–IV.

In addition, the results of both the hierarchical and the higher order CHC models demonstrated that the Similarities and Word Reasoning scores did not provide fluid reasoning (*Gf*) measures (Grégoire, 2009; Keith et al., 2006; Lecerf et al., 2010). Consistent with previous studies, the present results showed that the Block Design score loaded only on visual processing (*Gv*) and was not a measure of *Gf* (Lecerf et al., 2010). BSEM also indicated, unlike previous ML-CFA analysis, that the Block Design score was not a measure of processing speed (*Gs*). As concerns the Picture Completion score, our previous ML-CFA studies suggested that it primarily measures *Gc* and to a lesser degree *Gv*. In contrast, results with BSEM suggested a simpler interpretation of the Picture Completion score: This subtest measured only *Gv* and not *Gc*. Next, results indicated that the Matrix Reasoning subtest score mainly measured fluid reasoning (*Gf*), and had no substantive loading on *Gv*. This finding was not congruent with Carroll (1993), who suggested that fluid reasoning tests required visual processing. Consistent with our previous ML-CFA and in contrast to Keith et al. (2006), the Symbol Search score appeared to measure only processing speed (*Gs*); indeed, the cross-loading on *Gv* was not substantive. Finally, BSEM demonstrated that the Coding score loaded only on *Gs* and was not an indicator of *Gsm*. Altogether, besides the major loadings, no secondary interpretation of the subtest scores has been found to be backed up by the data. Thus, BSEM suggested simple and parsimonious interpretation of the subtest scores. Estimating all cross-loadings simultaneously did not lead to a more complicated loading pattern.

BSEM also provided some important additional information concerning the correlation between the first-order factors and the loadings on the second-order factor. Indeed, in contrast to the

results obtained with ML-CFA, results with BSEM indicated that the correlation between *g* and *Gf* was not 1 (or close to 1; .956 in the present investigation) but around .88. With classical ML-CFA estimation, some authors have suggested, statistical power could explain why it was not always possible to distinguish the general factor from fluid reasoning (Matzke, Dolan, & Molenaar, 2010). We suggest that inappropriate zero cross-loadings could also account for the unitary loading between *g* and *Gf*, because the correlations between first-order factors can be overestimated when using standard CFA estimators. Thus, in line with Matzke et al. and contrary to Gustafsson (1984), we argue that the equivalence of *g* and *Gf* in higher order models is better accounted for by statistical artifacts than the mere equivalence of the two constructs.

Third, the present study favored a direct hierarchical model, which was more adequate than a higher order model. In other words, the structure of the WISC–IV was best represented by five first-order CHC factors, plus a general intelligence factor. Consistent with previous studies, it was found that although subtest scores were aligned with their theoretically assumed factors (e.g., VCI or *Gc*), the *g* factor accounted for the largest portions of variance. Indeed, the variance explained by first-order factors was weak when the variance accounted for by the *g* factor was partialed out. For instance, 10 subtest scores exhibited a higher loading on *g* than on their respective first-order factor. Furthermore, results indicated that *g* explained the largest part of the variance of the *Gc*, *Gf,* and *Gv* subtest scores, in contrast to the *Gsm* and *Gs* factors. These findings are consistent with those reported by Watkins (2006) and Watkins, Wilson, Kotz, Carbone, and Babula (2006) for the U.S. WISC–IV, and they indicate more generally that published cognitive batteries may be overfactored (Canivez, 2008). In addition, this finding was consistent with Gignac's results (2008) and showed that the conceptualization of the general factor as a first-order breadth factor is preferable to that as a higher order superordinate factor: The relationship between the general factor and each subtest score does not seem to be fully mediated by the broad abilities. Nevertheless, we want to emphasize that this finding does not necessarily suggest that there is a single underlying mechanism for *g*. To the contrary, we consider that *g* is a complex construct defined by the interaction between several mechanisms (van der Maas et al., 2006). From a practical point of view, the present results indicated that interpretation of WISC–IV subtest scores should focus mainly on the Full Scale IQ (FSIQ) and that standard and/or CHC index scores do not necessarily provide additional and separate information. Remember that it has been shown that the general intelligence factor accounted for an important part of the FSIQ variance and that FSIQ was the best predictor of academic achievement regardless of index score variability (Watkins, Glutting, & Lei, 2007).

In conclusion, in the present paper, we used BSEM to analyze the structure of the French WISC–IV and to determine the nature of the constructs measured by each subtest scores. The results indicated that the BSEM approach to model specification and estimation performed better than the more classical and restrictive ML-CFA approach. The framework used in this study might be useful for other versions of the Wechsler scales and other intelligence tests. Although Bayesian methodology is used increasingly, to our knowledge it has not been frequently used with intelligence research and factor analysis of cognitive batteries, such as the WISC–IV. We argue that the Bayesian methodology is very efficient for incorporating both knowledge and uncertainty when analyzing psychological instruments.

References

Ackerman, P. L., & Lohman, D. F. (2006). Individual differences in cognitive functions. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 139–161). Mahwah, NJ: Erlbaum.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): What does the WAIS–IV measure? *Psychological Assessment, 22,* 121–130. doi:10.1037/a0017767

Canivez, G. L. (2008). Orthogonal higher order factor structure of the Stanford–Binet Intelligence Scales—Fifth Edition for children and adolescents. *School Psychology Quarterly, 23,* 533–541. doi:10.1037/a0012884

Canivez, G. L., & Watkins, M. W. (2010a). Exploratory and higher-order factor analyses of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV) adolescent subsample. *School Psychology Quarterly, 25,* 223–235. doi:10.1037/a0022046

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment, 22,* 827–836. doi:10.1037/a0020429

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

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30,* 429–452. doi:10.1207/s15327906mbr3003_6

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC–IV assessment*. New York, NY: Wiley.

Flanagan, D., Ortiz, S., & Alfonso, V. (2007). *Essentials of cross-battery assessment*. Hoboken, NJ: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–759.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472. doi:10.1214/ss/1177011136

Gignac, G. E. (2006). The WAIS–III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences, 27,* 73–86. doi:10.1027/1614-0001.27.2.73

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: *g* as superordinate or breadth factor? *Psychology Science Quarterly, 50,* 21–43.

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

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

Grégoire, J. (2009). *L'examen clinique de l'intelligence de l'enfant: Fondements et pratique du WISC–IV* [Clinical examination of child intelligence: Foundations and practice of WISC–IV]. Wavre, Belgium: Mardaga.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179–203. doi:10.1016/0160-2896(84)90008-4

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434. doi:10.1207/s15327906mbr2804_2

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

Kaufman, A. S., & Kaufman, N. L. (2008). *KABC batterie pour l'examen psychologique de l'enfant—Deuxième édition* [KABC Assessment Battery for Children—Second edition]. Paris, France: ECPA.

Keith, T. Z., Fine, J. G., Taub, G. E., 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,* 108–127.

Lecerf, T., Golay, P., Reverte, I., Senn, D., Favez, N., & Rossier, J. (2012). Scores composites CHC pour le WISC–IV: Normes francophones [CHC composite scores for the WISC–IV: French norms]. *Pratiques Psychologiques, 18,* 37–50. doi:10.1016/j.prps.2011.04.001

Lecerf, T., Rossier, J., Favez, N., Reverte, I., & Coleaux, L. (2010). The four- vs. alternative six-factor structure of the French WISC–IV. *Swiss Journal of Psychology, 69,* 221–232. doi:10.1024/1421-0185/a000026

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39,* 653–686. doi:10.1207/s15327906mbr3904_4

Link, W. A., & Eaton, M. J. (2011). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3,* 112–115. doi:10.1111/j.2041-210X.2011.00131.x

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504. doi:10.1037/0033-2909.111.3.490

MacEachern, S. N., & Berliner, L. M. (1994). Subsampling the Gibbs sampler. *American Statistician, 48,* 188–190.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471–491. doi:10.1037/a0019227

Matzke, D., Dolan, C. V., & Molenaar, D. (2010). The issue of power in the identification of "g" with lower-order factors. *Intelligence, 38,* 336–344. doi:10.1016/j.intell.2010.02.001

McCrae, R. R., Yamagata, S., Jang, K. L., Riemann, R., Ando, J., Ono, Y., . . . Spinath, F. M. (2008). Substance and artifact in the higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 95,* 442–455. doi:10.1037/0022-3514.95.2.442

McGrew, K. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37,* 1–10. doi:10.1016/j.intell.2008.08.004

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17,* 313–335. doi:10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell–Horn–Carroll–based assessment. *Psychology in the Schools, 47,* 621–634. doi:10.1002/pits.20495

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37–52. doi:10.1007/BF02294318

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61. doi:10.1007/BF02289209

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64,* 583–639. doi:10.1111/1467-9868.00353

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113,* 842–861. doi:10.1037/0033-295X.113.4.842

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

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

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., Glutting, J. J., & Lei, P. W. (2007). Validity of the Full Scale IQ when there is significant variability among WISC–III and WISC–IV factor scores. *Applied Neuropsychology, 14,* 13–20. doi:10.1080/09084280701280353

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,* 975–983. doi:10.1177/0013164406288168

Wechsler, D. (2005). *Manuel de l'Echelle d'intelligence de Wechsler pour enfants—Quatrième édition* [Manual for the Wechsler Intelligence Scale for Children—Fourth Edition]. Paris, France: ECPA.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III.* Itasca, IL: Riverside.

Yuan, Y., & MacKinnon, D. (2009). Bayesian mediation analysis. *Psychological Methods, 14,* 301–322. doi:10.1037/a0016972

Received February 28, 2012
Revision received September 21, 2012
Accepted September 28, 2012 ■