Assessment
2016, Vol. 23(6) 698–706
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191115595070
asm.sagepub.com
SAGE

Modeling General and Specific Abilities: Evaluation of Bifactor Models for the WJ-III

Dennis J. McFarland[1]

Abstract

The present study examined issues related to structural modeling of abilities by the use of simulated data as well as analysis of the standardization data from the Woodcock–Johnson-III. In both cases, results were evaluated with cross-validation. Simulation results showed that cross-validation with an independent data set was more successful in identifying the model that was used to generate test scores than were several fit indices. Analysis of the Woodcock–Johnson-III standardization data with cross-validation showed that bifactor models provided better fit than hierarchical or correlated factor models. This was true considering both fit indices and cross-validation. General and specific factors shared a considerable amount of variance as evaluated by using the bifactor models to partition variance. The results of the present study suggest that there is a certain degree of ambiguity in determining the exact amount of covariance in test performance accounted for by general and specific factors. This calls in to question the practice of adjusting or controlling for general abilities when evaluating measures of specific abilities. Evidence for the validity of a construct should not be limited to factor analysis of tests purported to measure that construct.

Keywords

WJ-III, bifactor models, cross-validation

Test batteries designed to assess specific cognitive abilities are often used in the identification of learning disabilities (e.g., Berninger, Abbott, Vermeulen, & Fulton, 2006; Floyd, Keith, Taub, & McGrew, 2007). Indeed, some researchers suggest that full-scale scores add little predictive power beyond that provided by more specific factors (e.g., Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001). Measures of achievement can be predicted by the results of factor analysis that group scales together that assess specific cognitive functions (Evans, Floyd, McGrew, & Leforgee, 2001). The Woodcock–Johnson (WJ-III) is one of the more common test battery used for this purpose (McGrew & Wendling, 2010).

In contrast to an emphasis on specific factors, several investigators have suggested that test battery interpretation should be made primarily on the basis of full-scale scores (e.g., Beaujean, Parkin, & Parker, 2014; Canivez, 2013; Golay & Lecerf, 2011; Nelson, Canivez, & Watkins, 2013; Watkins & Beaujean, 2014). These recommendations are based on the results of factor analysis studies that show that the general factor, or *g*, accounts for much more of the common variance than domain-specific factors.

There are a number of alternative approaches to modeling the covariance between tests of cognitive abilities. General and specific cognitive ability factors have often been conceptualized in terms of higher order models (e.g., Carroll, 1993). Higher order models treat the effect of *g* on test performance as being mediated by specific factors (Gignac, 2008). More recently, bifactor models that treat *g* and specific factors as independent determinants of behavior have gained popularity (Reise, 2012). Bifactor models have been shown to provide better fit of the covariance in cognitive test performance than higher order models (Beaujean et al., 2014; Gignac & Watkins, 2013; McFarland, 2013; Valerius & Sparfeldt, 2014). In addition, due to the independence of general and specific factors, the relative contribution of each to test performance can more readily be evaluated.

In comparing higher order models with bifactor models, Beaujean et al. (2014) state that when specific cognitive abilities are of concern, “the factor model used to represent abilities makes a great deal of difference, both conceptually and statistically” (p. 791). One reason the factor model is

[1]New York State Department of Health, Albany, NY, USA

Corresponding Author:
Dennis J. McFarland, Wadsworth Center, New York State Department of Health, P.O. Box 509, Empire State Plaza, Albany, NY 12201-0509, USA.
Email: dennis.mcfarland@health.ny.gov

important is that general and specific abilities are confounded in higher order models in which the effects of the general factor are mediated by the specific factors. The factor model is also of importance when researchers control for *g* when evaluating the impact of specific abilities (e.g., Vugs, Cuperus, Hendriks, & Verhoeven, 2013). Likewise, as already discussed, some researchers suggest that test batteries should be interpreted primarily in terms of *g*. However, others suggest that *g* is of little relevance to the diagnosis of specific learning disabilities (e.g., Hale et al., 2001). The relative importance of general and specific factors is related to the relative amount of variance accounted for by *g* and specific factors, which is model-dependent.

Murray and Johnson (2013) have questioned the wisdom of comparing bifactor and higher order models on the basis of model fit alone. Based on both logic and simulations, they suggest that there is an inherent statistical bias that favors the bifactor model. Higher order models can be conceptualized as models in which the ratio of the weights between any given specific factor and *g* is constant (Reise, 2012). In contrast, bifactor models do not have this constraint and hence have more degrees of freedom with which to fit the data. As a consequence, bifactor models are more complex than higher order models and are thus more prone to overfitting (Cudeck & Henly, 1991).

Overfitting occurs when model parameters account for chance characteristics of a sample rather than the underlying relationships they are intended to model. In statistics and machine learning, the problem of overfitting has frequently been dealt with by the use of cross-validation (Arlot & Celisse, 2010; Brown, 2000; Mosier, 1951). Cross-validation has also been recommended for covariance modeling, although it is rarely used (Browne & Cudeck, 1993; MacCallum, Roznowski, & Necowitz, 1992). With cross-validation, the model parameters are estimated with one sample (i.e., the training sample) and then these fixed parameter estimates are generalized to an independent sample. Overfitting is thus not an issue in assessing model fit in the second (i.e., the test) sample to which the model parameters were generalized. Anderson and Gerbing (1988) describe this procedure as "the quintessential confirmatory analysis" (p. 412).

The present study examined several issues related to structural modeling of abilities by the use of simulated data as well as analysis of the standardization data from the WJ-III (McGrew & Woodcock, 2001). The WJ-III standardization data were selected since it is a popular test battery designed to assess both general and specific abilities (McGrew & Woodcock, 2001), reported to have an age-invariant factor structure (Taub & McGrew, 2004) and has been previously examined with alternative factor models (e.g., Dombrowski & Watkins, 2013; McGrew & Woodcock, 2001). For both simulated and empirical data, results were evaluated with cross-validation. Several issues were evaluated. Simulation results evaluated whether cross-validation dealt with the issue of overfitting with bifactor models. Analysis of the WJ-III standardization data with cross-validation evaluated the relationship between general and specific factors.

Method

Simulations

All simulations were done in SAS with a C++ program used to organize the data. The basic model for the *k*th score on the *i*th test was

$$t_{ik} = \sum \left(w_{ij} \cdot a_{jk}\right) + e_i \qquad (1)$$

where a_{jk} is the magnitude of the *j*th ability for the *k*th observation, and e_i is a random test-specific term. The value of w_{ij} is the weight given a_{jk} on the *i*th test. The value of a_{ij} was unique to each individual within a test battery simulation and was drawn from the SAS normal distribution function. The value of a_{jk} represents the ability of an individual on some hypothetical trait (e.g., an individual's general intelligence or auditory processing ability), while the value of w_{ij} describes the role of these abilities in determining test performance (e.g., to what extent a test measures general intelligence or auditory processing).

Simulated test batteries consisted of 25 tests, simulated with one general factor and five specific factors. These simulations were based on either a higher order model or a bifactor model. In addition, the sample size (i.e., n = 200 or 2,000) and the amount of test-specific variance (i.e., the value of e_i being either 2× or 4× the value of the SAS normal distribution function, often referred to as error) were varied. Each of these conditions was simulated 12 times with different random values of w_{ij} for each of the multiple test battery simulations to extend the generality of findings. Each value of w_{ij} was unique to a single test battery simulation, and was drawn from the SAS uniform distribution function. Use of a uniform distribution insures that all abilities function in a similar manner within a given test battery simulation (i.e., if a given ability has positive effects on one test it would be expected to have positive effects on other tests). This is a boundary condition for all of the simulations conducted in the present study and was more extensively investigated by McFarland (2012). These simulations were done with SAS (2010) and the resulting data were then analyzed with the SAS CALIS procedure.

Participant Data

This study used the data reported for the standardization sample of the WJ-III (McGrew & Woodcock, 2001). Two samples of subjects were constructed consisting of data for individuals between 14 and 19 years of age (Table d-5), and

Table 1. Summary of Average Correlations Produced in Each Simulations.

Error	*n*	Model	Mean *r*	Minimum *r*	Maximum *r*
2×	200	Higher order	0.503	0.257	0.796
2×	2,000	Higher order	0.461	0.264	0.768
4×	200	Higher order	0.229	−0.024	0.498
4×	2,000	Higher order	0.224	0.084	0.460
2×	200	Bifactor	0.305	0.036	0.648
2×	2,000	Bifactor	0.297	0.126	0.610
4×	200	Bifactor	0.126	−0.104	0.374
4×	2,000	Bifactor	0.122	0.021	0.293

Note. Averages were computed by first taking Fisher's *z* transformation, averaging, and then converting the average back to Pearson's *r* values. Model refers to the model used to simulate test scores.

individuals between 20 and 39 years of age (Table d-6). Two samples of tests were constructed. The first assigned WJ-III cognitive tests to the factors that they were associated with in Table 2-2 of the WJ-III technical manual (McGrew & Woodcock, 2001) and hereafter referred to as the WJ-III model. The second consisted of the model identified in Table 7 of Dombrowski and Watkins (2013) and hereafter referred to as the Dombrowski model. The Dombrowski model is based on a subset of the combined WJ-III cognitive and achievement scales that had large loadings on a five-factor solution for the WJ-III standardization data from 14 to 19 year olds.

Analyses

Three models were compared for each of the simulated test conditions and each of the participant samples. These were a higher order model, a correlated factors model, and a bifactor model. For simulated data, the structure of the specific factors used in the models were identical to those used to generate the data. For participant data, the structure of the specific factors were those in the WJ-III (McGrew & Woodcock, 2001) and Dombrowski models (Dombrowski & Watkins, 2013).

All analyses were done with the SAS CALIS procedure (SAS, 2010) using default settings. All latent factors were set equal to 1 (except for those defined by a hierarchical structure) as recommended by Anderson and Gerbing (1988). Fit indices included chi-squared (χ^2), Bentler's comparative fit index, the standardized root mean square error, the standardized root mean square residual, and the Akaike information criterion. These indices were selected so as to provide a comparison with prior studies. These indices differ in how they deal with the trade-off between model accuracy and complexity. The fact that there has been a proliferation of such indices attests to the difficulty of equating accuracy and complexity. This problem of model evaluation is also dealt with by the use of cross-validation (Brown, 2000; Mosier, 1951). Cross-validation in the present study used the loadings for each factor on each subtest as estimated from the data sets used to estimate parameters (training set) and applied to independent samples (test sets). Only the scale-specific effects (error) were estimated in the evaluation of models with cross-validation, as is also the case with the NULL model, which was also included. This approach is the fixed-structure strategy described by MacCallum, Roznowski, Mar, and Reith (1994). With this approach, the number of estimated parameters (i.e., model complexity) is identical in fixed-structure, cross-validated models.

Results

Average correlations from the various simulation conditions are shown in Table 1. The results of simulations using a higher order model to generate test scores are shown in Table 2. Single-factor analyses of variance were used to compare the fit of each model (higher order, correlated factor, and bifactor) for each fit index within each simulation condition. Tukey's studentized range test was applied to all significant effects. The χ^2 values for the cross-validated data (χ^2 residual from generalized weights) resulted in significantly lower values (i.e., better fit) for the bifactor model with all four simulation conditions. The use of the χ^2 residual of the test sample using weights derived from the training sample is essentially the cross-validation index discussed by MacCallum et al. (1994, their Formula 2). Each of the other indices indicated that either the correlated factor or the bifactor model fit the data significantly better, with the exception of comparative fit index in the 2×-2,000 condition, for which the models did not differ significantly. Thus, like Murray and Johnson (2013), the present simulations found that these fit indices were biased toward the bifactor model. However, cross-validation consistently favored the true higher order model in these comparisons.

The results of simulations using a bifactor model to generate test scores are shown in Table 3. All of the fit indices indicated that the bifactor model used to generate the data was also the best fitting model, with the exception of AIG

Table 2. Summary of Mean Values for the Various Fit indices for Data Generated With a Hierarchical Model.

Error	*n*	Model	χ^2 train	CFI	RMSEA	SRMR	AIC	χ^2 test
2×	200	Hier	302.084^{c}	0.9884^{b}	0.0247^{b}	0.0383^{c}	422.084^{b}	322.971^{a}
2×	200	Corr	295.154^{b}	0.9904^{a}	0.0221^{a}	0.0357^{b}	415.154^{a}	342.828^{b}
2×	200	Bi-F	281.395^{a}	0.9898^{a}	0.0228^{ab}	0.0331^{a}	431.396^{c}	351.823^{b}
2×	2,000	Hier	285.306^{c}	0.9988	0.0057^{b}	0.0124^{c}	405.306^{b}	331.578^{a}
2×	2,000	Corr	278.021^{b}	0.9995	0.0043^{a}	0.0115^{b}	398.021^{a}	351.078^{b}
2×	2,000	Bi-F	263.956^{a}	0.9994	0.0049^{ab}	0.0106^{a}	413.956^{c}	362.820^{c}
4×	200	Hier	302.689^{c}	0.9612^{b}	0.0246^{b}	0.0518^{c}	422.689^{b}	323.878^{a}
4×	200	Corr	295.257^{b}	0.9676^{a}	0.0218^{a}	0.0506^{b}	415.257^{a}	343.534^{b}
4×	200	Bi-F	280.739^{a}	0.9677^{a}	0.0228^{a}	0.0486^{a}	430.739^{c}	355.523^{c}
4×	2,000	Hier	287.609^{c}	0.9969^{b}	0.0057^{b}	0.0166^{c}	407.609^{b}	332.497^{a}
4×	2,000	Corr	281.734^{b}	0.9977^{ab}	0.0050^{a}	0.0163^{b}	401.734^{a}	350.582^{b}
4×	2,000	Bi-F	264.606^{a}	0.9979^{a}	0.0049^{a}	0.0153^{a}	414.606^{c}	365.109^{c}

Note. CFI = comparative fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error; AIG = akaike information criterion; Bi-F = bifactor; Corr = correlated; Hier = hierarchical. Means with the same superscript are not different.

Table 3. Summary of Mean Values for the Various Fit Indices for Data Generated With a Bifactor Model.

Error	*n*	Model	χ^2 train	CFI	RMSEA	SRMR	AIC	χ^2 test
2×	200	Hier	345.548^{b}	0.9616^{b}	0.0378^{b}	0.0600^{b}	465.548^{b}	368.773^{a}
2×	200	Corr	338.480^{b}	0.9649^{b}	0.0358^{b}	0.0572^{b}	458.480^{b}	389.167^{b}
2×	200	Bi-F	279.870^{a}	0.9846^{a}	0.0221^{a}	0.0440^{a}	429.870^{a}	353.653^{a}
2×	2,000	Hier	721.82^{b}	0.9780^{b}	0.0291^{b}	0.0378^{b}	841.82^{b}	776.93^{b}
2×	2,000	Corr	714.78^{b}	0.9784^{b}	0.0288^{b}	0.0373^{b}	834.78^{b}	796.41^{b}
2×	2,000	Bi-F	264.64^{a}	0.9990^{a}	0.0052^{a}	0.0139^{a}	414.64^{a}	365.57^{a}
4×	200	Hier	309.830^{c}	0.9097^{c}	0.0276^{b}	0.0602^{c}	429.830^{b}	332.325^{a}
4×	200	Corr	302.310^{b}	0.9229^{b}	0.0252^{ab}	0.0586^{b}	422.310^{a}	352.342^{b}
4×	200	Bi-F	279.447^{a}	0.9380^{a}	0.0222^{a}	0.0553^{a}	429.448^{b}	356.843^{b}
4×	2,000	Hier	354.072^{b}	0.9812^{b}	0.0129^{b}	0.0223^{b}	474.072^{b}	405.423^{b}
4×	2,000	Corr	348.355^{b}	0.9823^{b}	0.0124^{b}	0.0220^{b}	468.355^{b}	423.589^{b}
4×	2,000	Bi-F	265.045^{a}	0.9959^{a}	0.0054^{a}	0.0174^{a}	415.045^{a}	367.185^{a}

Note. CFI = comparative fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error; AIG = akaike information criterion; Bi-F = bifactor; Corr = correlated; Hier = hierarchical. Means with the same superscript are not different.

and the cross-validated χ^2 in the 4× error, $n = 200$ condition. Thus, when the data were simulated with a bifactor model the same bifactor model provided a better fit except when error was high and the sample size was low. This finding is consistent with the observations of Cudeck and Henly (1991) that fit indices may be biased toward simple models when sample sizes are small, even when complex models may provide a better fit in the entire population. Overall, cross-validation with large populations would appear to be advantageous when practical.

Table 4 shows the average of the results of correlating the general and specific factors used to generate the simulated bifactor data with the factor scores extracted with the structural models based on the default maximum-likelihood method. As can be seen in Table 4, the largest correlations are on the diagonal, as would be expected if the extracted factors were related to the generating factors. However, there is considerable off-diagonal variance, particularly for the extracted general factor and the generating specific factors. This indicates that the extracted factors contain a considerable amount of variance related to all of the generating sources of variance. This is unlikely to be due to chance as the sample size for these simulations are large (i.e., $n = 2,000$). Thus, although the source factors used to generate the simulated data were orthogonal, the resulting model factors were composites of multiple sources.

The results of evaluating variations on the structural model presented in Table 2-2 of the WJ-III test manual on two samples from the standardization data (McGrew & Woodcock, 2001) are presented in Table 5. The original model was compared with models that used the same specific factors but dealt with the general factor either by allowing the specific factors to be correlated or by use of a bifactor model. As can be seen in Table 4, the bifactor model

Table 4. Average Correlations of Generative Variables (Sg and S1 Through S5) With Factor Scores Obtained From the Structural Models (Fg and F1 Through F5). (Bifactor, *n* = 2,000, *e* = 4×).

	Sg	S1	S2	S3	S4	S5
Fg	0.8120	0.1874	0.2067	0.2352	0.2206	0.1597
F1	0.1285	0.5859	0.0973	0.0994	0.0900	0.0479
F2	0.1607	0.0772	0.5776	0.0823	0.1010	0.0632
F3	0.1454	0.0863	0.0972	0.5493	0.0939	0.0685
F4	0.1667	0.0844	0.0841	0.0892	0.5882	0.0798
F5	0.1185	0.0525	0.1078	0.0830	0.0651	0.6334

Table 5. WJ-III Models Trained on 16- to 19-Year Data and Generalized to 20- to 39-Year Data.

Model	*df*	χ^2	GFI	AGFI	RMSEA	Generalized χ^2	Generalized GFI
WJ-III	155	1132.1320	0.9210	0.8930	0.0697	1486.4009	0.8776
WJ-III correlated	148	1026.5278	0.9282	0.8981	0.0676	1326.4369	0.8917
WJ-III bifactor	149	809.3914	0.9405	0.9161	0.0548	1278.9827	0.8921

Note. WJ-III = Woodcock–Johnson-III; *df* = degrees of freedom; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; RMSEA = root mean square error. WJ-III model is from manual.

Table 6. Variations on the Model Derived From Table 5 in Dombrowski and Watkins (2013) Trained on 16- to 19-Year Data and Generalized to 20- to 39-Year Data From the WJ-III Standardization Sample.

Model	*df*	χ^2	GFI	AGFI	RMSEA	Generalized χ^2	Generalized GFI
Dombrowski higher order	264	3866.1714	0.8098	0.7659	0.0997	4106.5324	0.7684
Dombrowski correlated	264	3823.8530	0.8131	0.7699	0.1019	4123.2148	0.7641
Dombrowski bifactor	249	2641.0400	0.8524	0.8073	0.0860	2795.9331	0.8216

Note. WJ-III = Woodcock–Johnson-III; *df* = degrees of freedom; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; RMSEA = root mean square error. WJ-III model is from manual.

provided the best fit considering all fit indices as well as with cross-validation.

The results of evaluating variations on the model derived from Table 7 in Dombrowski and Watkins (2013) are presented in Table 6. Again, the original model was compared with models that used the same specific factors but dealt with the general factor either by allowing the specific factors to be correlated or by use of a bifactor model. As can be seen in Table 4, the bifactor model provided the best fit for all fit indices and for the cross-validation data.

Table 7 presents the bifactor models from Tables 5 and 6 along with models that consist of either just the general factor or just the five specific factors. Based on a comparison with the null model, the amount of shared and unique variance accounted for by each model is also presented in Table 7. For the models based on Table 2-2 of the WJ-III manual, the five specific factors alone account for more variance than the single general factor alone. This is true for both training and test sets. For the models based on Table 5 of Dombrowski and Watkins (2013), the single general factor alone accounts for more variance than the five specific factors alone. This is true for both training and test sets. For both training and test data sets, there is a large component that is shared. Indeed, the shared component is larger than either of the unique components in both test data sets. This is analogous to regression problems where correlated predictors result in a situation where assigning the shared variance to specific variables is ambiguous and requires assumptions about causality (Overall & Spiegel, 1969).

Discussion

Simulation results showed that cross-validation with an independent data set was more successful in identifying the model that was used to generate test scores than were several fit indices. An exception is the case for complex models that are evaluated with small samples. Evaluation of two different partitions of the WJ-III test battery showed that bifactor models provided better fit than hierarchical or correlated factor models. This was true considering both fit indices and cross-validation. General and specific factors

Table 7. Bifactor Decomposition.

Model	WJ-III train	WJ-III test	Dombrowski train	Dombrowski test
NULL	12071.2177	12951.4009	25345.9122	23653.0977
Bifactor	809.3914	1278.9827	2641.0400	2795.9331
One General	4150.6537	4596.0680	8937.6412	8881.5995
Five Specific	4616.1208	5145.9108	6827.5445	6689.7358
General alone	3806.7294	3866.9281	4186.5045	3893.8027
Specific alone	3341.2623	3317.0853	6296.6012	6085.6664
Shared	2495.0518	4488.4048	12221.7665	10877.6955

Note. WJ-III = Woodcock–Johnson-III. One General is the difference between the full bifactor model and the general factor omitted. Five Specific is the difference between the full bifactor model and the 5 specific factors omitted. Shared is the difference between the NULL model and the bifactor model less the independent effects of the 1 and 5 factors.

shared a considerable amount of variance as evaluated by using the bifactor models to partition variance.

In the present study, models were compared in terms of generalization of the parameters estimated from the 14- to 19-year old samples to the 20- to 39-year old sample. This represents what Mosier (1951) referred to as validity generalization since the new samples represent different populations rather than simply an additional sample from the same population. The generalization of the bifactor models to data from a new population provides strong evidence that the superior fit of these models is not due to overfitting. In addition, models that generalize more broadly are arguably more useful.

Murray and Johnson (2013) have suggested that there is an inherent statistical bias that favors the bifactor model over hierarchical models when considering only fit indices. They base this conclusion in part on the fact that bifactor models have more degrees of freedom with which to fit "unmodeled complexity." Murray and Johnson (2013) also present the results of simulations that support this conclusion. However, they only evaluated simulations based on data generated by hierarchical models. Murray and Johnson (2013) suggest that unmodeled complexity is due to the fact that psychometric models of human cognitive ability represent simplifications of its true structure. They further suggest that although this unmodeled complexity may be large, any one aspect of it may be too small to merit inclusion in the model, or it may be due to sampling fluctuations that give rise to chance intercorrelations. The issues discussed by Murray and Johnson (2013) involve both theoretical considerations and statistical methodology. The theoretical issues are debatable, such as whether models of human abilities should be viewed as "correct" or simply useful (e.g., McFarland, 2014). The statistical issue of overfitting due to capitalization of chance variance can be dealt with by cross-validation. As shown in the present series of simulations, given a large sample size, cross-validation can identify the best model. Indeed, with small sample sizes there may at times be a bias against complexity (Cudeck & Henly, 1991). This is illustrated by the results from the present study where data generated using a bifactor model was better accounted for by a hierarchical model in simulations with the smaller sample size.

The view that a single general factor is the major contribution to cognitive test scores is based in part on the results of factor analysis (Canivez, 2013; Dombrowski & Watkins, 2013). As shown in the present study, based on the bifactor model, there is considerable overlap in the variance that can be accounted for by a general factor and several specific factors. Assigning this common variance to either the general factor or the specific factors requires adoption of some model that has causal implications. The issue of common predicted variance in structural modeling is analogous to the problem of collinear predictors in multiple regression. Hale, Fiorello, Kavanagh, Holdnack, and Aloe (2007) have shown that the order of entry of correlated variables in hierarchical regression has a marked effect on the assignment of variance to general and specific factors in predicting achievement from tests of abilities. They suggest entering specific factors first. Based on the principal of parsimony, Canivez (2013) rejected this suggestion and considered an alternative that assigns the maximum amount of shared variance to a general factor. However, parsimony is not the only issue to consider since more complex models may be appropriate if they have greater utility. Partitioning the variance predicted by correlated variables is problematic, particularly when, as in the Hale et al. (2007) case, the Wechsler Intelligence Scale for Children–Fourth edition general factor is a weighted composite of the four specific factors. One advantage of the bifactor model in such cases is that the general and specific factors can be modeled with factor correlations set to zero (as was done in the present study). Thus, use of factor scores derived from bifactor models to predict achievement should be less affected by problems of collinearity.

It is a common practice to adjust or control for general abilities when evaluating measures of specific abilities. For example, Vugs et al. (2013) included only studies using subjects with normal nonverbal intelligence in a meta-analysis of the effects of visual working memory on specific language impairment. Vaessen, Gerretsen, and Blomert (2009) included verbal IQ scores in a hierarchical regression step prior to evaluating the effects of phonological awareness and rapid automatized naming on reading and spelling. Procedures such as these are based on the assumption that the variance accounted for by tests of specific abilities should make a contribution beyond that accounted for by general abilities. However, this view is by no means universal (e.g., Cahan, Fono, & Nirel, 2012; Siegel, 2003).

Using simulations, McFarland (2014) showed that cases selected on the basis of low true scores on a specific factor could appear to differ mainly on a general factor derived from factor analysis when used in hierarchical regression. This is due to the fact that the criterion for optimizing the weights of the first principal component is to maximize the amount of covariance accounted for, rather than to detect the "true" structure of individual differences. As a result, the first principal component may actually contain covariance that was generated by specific factors (McFarland, 2014). This sort of problem in the interpretation of factor analysis has been known for a long time (e.g., Overall, 1964). The results presented in Table 4 of the current study likewise suggest that factors may contain heterogeneous sources of variance. While any such result depends on methodology used for factor extraction, they do indicate cause for caution in the interpretation of factors.

The results of simulations do not prove that certain models are true, they only point to possibilities. Simulations based on the premise that most of the variance in test scores is due to a general factor show that principal components produces the correct result (e.g., Velicer & Fava, 1998). Simulations based on other assumptions show that principal components may not always produce a good result (e.g., McFarland, 2012, 2014). The present simulation results also provide some useful suggestions concerning methodology. They show that simple models may have an advantage over more complex models when sample sizes are small. This may be the case even when the more complex model is in fact the true model (i.e., results for the large error, small sample size condition in Table 3). The present simulation results also show that cross-validation to new data provides a useful way of evaluating alternative models. Although cross-validation has been recommended for use in structural equation modeling (Browne & Cudeck, 1993; MacCallum et al., 1992) and is commonly used in the machine learning community (Efron & Gong, 1983), it has received very little attention in the psychometric literature. There has been a plethora of suggestions for the use of various fit indices that adjust for model complexity (e.g., Yuan & Bentler, 1998). Fit indices suffer from the problem of determining the tradeoff between goodness-of-fit and complexity. Cross-validation provides a solution to this problem, provided the sample size is sufficient. Furthermore, cross-validation provides a means of determining generalizability to samples that differ from the original training sample.

The results of the present study suggest that there is a certain degree of ambiguity in determining the exact amount of covariance in test performance accounted for by general and specific factors. Since general and specific factors are independent in bifactor models, it was possible to evaluate the unique contribution of each (Table 7). The results show that there is considerable covariance that is not unique to either. This issue cannot be resolved solely by analysis of the covariance structure of test performance. Other sources of information are required, such as that provided by genetics, neurophysiology, and utility in predicting prognosis and response to treatment.

Utility in predicting prognosis or treatment response might vary with the nature of the constructs that serve as predictors as well as those that are predicted. For example, if the purpose is to predict prognosis or response to treatment for a specific learning problem, then a construct related to a specific ability might be most appropriate. For example, Taub, Floyd, Keith, and McGrew (2008) examined the prediction of mathematics achievement by WJ-III cognitive abilities using covariance structural modeling. After sequentially eliminating nonsignificant paths between abilities and achievement, they found that fluid reasoning, crystalized intelligence, and processing speed remained as predictors. In contrast, prediction of prognosis related to broad areas of functioning or for which there are a limited number of validity studies available might best employ a more general construct. For example, Strenze (2007) conducted a meta-analysis of studies relating intelligence to socioeconomic success and found a positive association between the two. In this case, the heterogeneity of indicators might make it difficult to model specific abilities so that a general ability construct might be more useful. This could be so even if the general construct actually is a composite of several specific abilities.

The ultimate test of utility should be actual empirical findings on the validity of the constructs in question. Evidence for the validity of a construct should not be limited to factor analysis of tests purported to measure that construct as this evidence is model-dependent. Assertions that interpretation of cognitive test batteries should be made primarily on the basis of full-scale scores (Canivez, 2013; Watkins & Beaujean, 2014) are based on specific measurement models that have not been compared with other possible alternatives. Furthermore, evaluating the validity of constructs requires relating these to measures of the specific target attributes that are to be predicted. Thus, sweeping generalizations about the use of test batteries such as the WJ-III are probably not warranted. Rather consideration of evidence for each specific application is probably more advisable.

Acknowledgments

I would like to thank Loretta Malta for her helpful comments on an earlier version of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79.

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll factor models: Differences between bifactor and higher order models in predicting language achievement. *Psychological Assessment*, *26*, 789-805.

Berninger, V. W., Abbott, R. D., Vermeulen, K., & Fulton, C. M. (2006). Paths to reading comprehension in at-risk second-grade readers. *Journal of Learning Disabilities*, *39*, 334-351.

Brown, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108-132.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation model* (pp. 136-162). Newbury Park, CA: Sage.

Cahan, S., Fono, D., & Nirel, R. (2012). The regression-based definition of learning disability: A critical appraisal. *Journal of Learning Disabilities*, *45*, 170-178.

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *Oxford handbook of child psychological assessment* (pp. 84-112). New York, NY: Oxford University Press.

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

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512-519.

Dombrowski, S. C., & Watkins, M. W. (2013). Exploratory and higher order factor analysis of the WJ-III full test battery: A school-aged analysis. *Psychological Assessment*, *25*, 442-455.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2001). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review*, *31*, 246-262.

Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S. (2007). Cattell–Horn–Carroll cognitive abilities and their effects on reading decoding skills: *g* has indirect effects, more specific abilities have direct effects. *School Psychology Quarterly*, *22*, 200-233.

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

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*, 639-662.

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.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaither, R. A. (2001). WISC-III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly*, *16*, 31-55.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Holdnack, J. A., & Aloe, A. M. (2007). Is the demise of IQ interpretation justified? A response to special issue authors. *Applied Neuropsychology*, *14*, 37-51.

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*, 1-32.

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.

McFarland, D. J. (2012). A single g factor is not necessary to simulate positive correlations between cognitive tests. *Journal of Clinical and Experimental Neuropsychology*, *34*, 378-384.

McFarland, D. J. (2013). Modeling individual subtests of the WAIS IV with multiple latent factors. *PLOS One*, *8*, e74980.

McFarland, D. J. (2014). Simulating the effects of common and specific abilities on test performance: An evaluation of factor analysis. *Journal of Speech, Language, and Hearing Research*, *57*, 1919-1928.

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools*, *47*, 651-675.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual*. Itasca, IL: Riverside.

Mosier, C. I. (1951). The need and means of cross-validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5-11.

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.

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.

Overall, J. E. (1964). Note on the scientific status of factors. *Psychological Bulletin*, *61*, 270-276.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *72*, 311-322.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667-696.

SAS. (2010). *SAS/STAT 9.22 user's guide*. Cary, NC: SAS Institute.

Siegel, L. S. (2003). IQ-discrepancy definitions and the diagnosis of LD: Introduction to the special issue. *Journal of Learning Disabilities*, *36*, 2-3.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*, 401-426.

Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*, 187-198.

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly*, *19*, 72-87.

Vaessen, A., Gerretsen, P., & Blomert, L. (2009). Naming problems do not reflect a second independent core deficit in dyslexia: Double deficits explored. *Journal of Experimental Child Psychology*, *103*, 202-221.

Valerius, S., & Sparfeldt, J. R. (2014). Consistent *g*- as well as consistent verbal-, numerical- and figural-factors in nested models? Confirmatory factor analysis using three test batteries. *Intelligence*, *44*, 120-133.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231-251.

Vugs, B., Cuperus, J., Hendriks, M., & Verhoeven, L. (2013). Visuospatial working memory in specific language impairment: A meta-analysis. *Research in Developmental Disabilities*, *34*, 2586-2597.

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition. *School Psychology Quarterly*, *29*, 52-65.

Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289-309.