Intelligence 51 (2015) 79–97

Contents lists available at ScienceDirect

Intelligence

Testing Spearman's hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data☆,☆☆

Craig L. Frisby [a,*], A. Alexander Beaujean [b]

[a] Department of Educational, School, and Counseling Psychology, University of Missouri, United States
[b] Educational Psychology Department, Baylor University, United States

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form 13 April 2015
Accepted 16 April 2015
Available online 8 June 2015

Keywords:
Spearman's hypothesis
Bi-factor model
Wechsler Adult Intelligence Scale-Fourth Edition
Wechsler Memory Scale-Fourth Edition

ABSTRACT

Spearman's hypothesis (SH) is a phrase coined by Arthur Jensen, which posits that the size of Black–White mean differences across a group of diverse mental tests is a positive function of each test's loading onto the general intelligence (*g*) factor. Initially, a correlated vector (CV) approach was used to examine SH, where the results typically confirmed that the magnitude of *g* loadings were positively correlated with the size of mean group differences in the observed test scores. The CV approach has been heavily criticized by scholars who have argued that a more precise method for examining SH can be better investigated using a multi-group confirmatory factor analysis (MG-CFA). Studies of SH using MG-CFA have been much more equivocal, with results not clearly confirming nor disconfirming SH.

In the current study, we argue that a better method for extracting *g* in both the CV and MG-CFA approaches is to use a bi-factor model. Because non-*g* factors extracted from a bi-factor approach are independent of *g*, the bi-factor model allows for a robust examination of the influence of *g* and non-*g* factors on group differences on mental test scores. Using co-normed standardization data from the Wechsler Adult Intelligence Scale-Fourth Edition and the Wechsler Memory Scale-Fourth Edition, we examined SH using both CV and MG-CFA procedures. We found support for the weak form of SH in both methods, which suggests that both *g* and non-*g* factors were involved in the observed mean score differences between Black and White adults.

1. Introduction

Differences between racial, ethnic, and socioeconomic groups in mean scores on general cognitive ability tests are well-established (Gottfredson, 2005; Rushton & Jensen, 2005). The magnitude of these differences, however, varies as a function of the type of cognitive skills being measured, with tests that more strongly related to general intelligence (*g*) exhibiting larger group differences (Jensen, 1998). Efforts to explain these patterns in the magnitude of group performance differences range from non-empirical speculations to those grounded in theory and appropriate empirical procedures for testing hypotheses.

1.1. Speculative Explanations

Speculative explanations simply proffer plausible, but ad hoc, rationales for why a particular group obtains lower mean scores than another group. These explanations are not tied to a coherent, data-based theory. As one example, "cultural differences" is often evoked as a global, all-purpose explanation for differing performance patterns among population subgroups. This global explanation typically takes two forms. Some may

☆ Standardization data from the *Wechsler Adult Intelligence Scale, Fourth Edition* (WAIS-IV). Copyright © 2008 NCS Pearson, Inc. Used with permission. All rights reserved.
☆☆ Standardization data from the *Wechsler Memory Scale, Fourth Edition* (WAIS-IV). Copyright © 2009 NCS Pearson, Inc. Used with permission. All rights reserved.

* Corresponding author at: Department of Educational, School, and Counseling Psychology, 16 Hill Hall, Columbia, MO 65211, United States.
E-mail address: FrisbyCL@missouri.edu (C.L. Frisby).

http://dx.doi.org/10.1016/j.intell.2015.04.007
0160-2896/

argue that some subgroups, partly due to economic and social disadvantages/differences from the more affluent mainstream, are simply not exposed to certain academic stimuli as is the case with more advantaged subgroups (Eells, 1951; Lupi & Ting Woo, 1989; White, 1984), and thereby lower scores are due to a presumed lack of exposure to tasks such as those found on cognitive tests (see *specificity doctrine*; Jensen, 1984). Others may argue that examinees from different racial/cultural groups display different "culturally idiosyncratic" psychological and/or stylistic patterns for interacting with test material, thereby depressing scores (see Helms, 1992, 1997).

Speculative explanations suffer from two major flaws. First and fundamentally, findings are explained only *after* they are observed. Testable hypotheses are not stated first before any data has been collected, which would allow for a rejection of the hypotheses based on patterns shown by the data. Second, these ad hoc explanations are infinitely malleable, adapting indiscriminately to the idiosyncratic characteristics of test items. As examples, Helms (1997) hypothesized that Black examinees may fail the Wechsler Intelligence Scales Arithmetic items because of substandard training in school, may fail Comprehension items due to "exposure to racism", and may fail Digit Symbol items because they are "uncomfortable with pencils as a tool" (p. 522).

1.2. Theory-based explanations: Spearman's hypothesis

Charles Spearman (1927) initially observed that race differences should be "most marked in just those [tests] which are known to be saturated with *g* [general intelligence]" (p. 379). Jensen (1980) later named this *Spearman's hypothesis* (SH). There are three levels of SH that Jensen (1998, 2001) called the strong form, weak form, and the contra hypothesis. The *strong form* posits that any observed race differences in test's mean scores are solely a function of *g*. The *weak form* posits that while race differences in test score means are mainly a function of *g*, lower-order factors or subtest specificities also contribute to the difference. The *contra hypothesis* holds that observed mean score differences are independent of *g*, being solely a function of lower-order factors or test specificity.

Support for SH has been borne out from numerous independent studies based on large child and adult samples (e.g., Jensen, 1985, 1998) and comprising many different psychometric tests, such as the Armed Forces Qualification Test (Nyborg & Jensen, 2000), the Kaufman Assessment Battery for Children (Naglieri & Jensen, 1987), the Wechsler Intelligence Scale for Children-Revised (Jensen & Reynolds, 1982; Naglieri & Jensen, 1987; Rushton & Jensen, 2003), and tests for college/graduate school admissions, job selection, and the military (Roth, Bevier, Bobko, Switzer, & Tyler, 2001).

There has been some disagreement about interpreting the SH literature. Schönemann (1997) interpreted the literature as being supportive of the weak form of SH. In contrast, Rushton (2003) concluded that most studies supported the strong form of SH.[1] Summarizing his own work from 17 independent data sets that included scores from 149 different tests obtained on samples of 45,000 Black and 245,000 White examinees, Jensen (2001) found that the correlation between Black–White differences and *g* was between .57 and .62. More recently, Dragt (2010) performed a meta-analysis of SH studies and found an average correlation of .85 between *g* and mean group test score differences between Black and White respondents.

1.2.1. Interpretation of Spearman's hypothesis using tests of memory

While Jensen's (Jensen & Figueroa, 1975; Jensen & Osborne, 1979) initial interest in SH began with tests of memory, little work has been done examining Black–White differences in memory measures. What has been done is mostly incidental (i.e., one or two memory subtests in an intelligence test battery), but it tends to indicate both that memory tasks have smaller *g* loadings than other tasks on multi-test cognitive batteries and that Black–White differences in mean scores are either considerably reduced on such tests (Jensen, 1980, 1985) or that average score for the Black sample is higher than the average for the White sample (Jensen & Reynolds, 1982). For example, in one of the few studies that examined Black–White differences in a battery of memory tests, Mayfield and Reynolds (1997) found a consistent factor structure across both groups. The Black sample scored higher than the White sample on most of the memory tests, although the difference was small.

1.3. Empirical challenges to interpretations of Spearman's hypothesis (SH)

Helms-Lorenz, Van de Vijver, and Poortinga (2003) have argued that the constructs of cognitive complexity and verbal/cultural loading are confounded in attempts to properly interpret results from tests of SH. They administered two intelligence batteries and a computer-assisted elementary cognitive test battery to a large group of Dutch and second-generation migrant 6–12 year old children living in the Netherlands. In addition to using factor analysis to compute the subtests' *g* loadings, they gave all subtests two ordinal ratings of "cognitive complexity." One cognitive complexity rating was based on both Carroll's (1993) cognitive abilities model while the other corresponded to the minimal developmental level needed for successful accomplishment (Fischer, 1980). The cultural loading of subtest content was rated on an ordinal scale by psychology students, and another rating of each subtest's verbal loading was operationalized as the number of words in the subtest. The authors found that the size of group differences on the intelligence tests was better predicted by the "cultural" variables than by the cognitive complexity variables.

Although Helms-Lorenz et al. (2003) used an intriguing methodology for investigating the relationship between factor-analytically derived subtest *g* loadings and human ratings of subtest task characteristics, there are a number of unresolved issues that challenge their conclusions. The first problem concerns confusion in what Jensen (1998) called the "vehicles of *g*" versus the *g* construct itself (Jensen, 1998, p. 309). For example, cultural differences would not explain why a Forward Digit Span Task and a Backward Digit Span Task would show widely discrepant *g* loadings, despite similarities in the surface characteristics of these tests (particularly in their nonverbal content). In addition, the composition of the comparison

[1] Rushton (1998) proposed that the term *Jensen Effect* be used whenever there is a substantial correlation between *g* loadings and any other variable.

groups that Helms-Lorenz et al. used may play a role in their findings. Jensen (1998) wrote:

> Each test score reflects both the level of *g* and the properties of the vehicle of *g* (the latter being largely unrelated to *g*). One would predict, for example, that the *g* factor, which is highly and equally loaded in batteries of verbal and nonverbal tests when given to monolingual children, would have much smaller *g* loadings on the verbal tests (given in English) than on the nonverbal tests when that battery is given to bilingual children. For the bilingual group the verbal tests would reflect the degree of second-language acquisition more than they would reflect g (p. 310).

Although the groups studied in the Helms-Lorenz et al. (2003) research are reported to have been exposed to the same number of (age appropriate) years of Dutch education, they also state "there is evidence that substantial differences in knowledge of the Dutch lexicon between the majority-group pupils and migrant pupils remain throughout the primary school period, even for second-generation children" (pp. 14–15). In the majority of studies that have evaluated SH, the comparison groups are comprised of native-born participants (e.g., American blacks and whites). In these studies, the comparison groups are more "culturally homogeneous" than those in the Helms-Lorenz et al. (2003) study where the migrant students' parents were born in at least five different countries.

1.4. Methods used to test for Spearman's hypothesis

There are two common methods currently employed to assess SH: correlated vector (CV) analysis, and multi-group confirmatory factor analysis (MG-CFA).

1.4.1. Correlated vector method

A correlated vector (CV) analysis attempts to explain variability in the magnitude of group differences on various tests (or subtests) by correlating the *g* loading of the tests with the size of group differences in mean scores on the same tests. A CV analysis typically involves the following steps: (a) conduct an exploratory factor analysis (EFA) of the tests in representative samples of the different comparison groups, separately; (b) estimate the similarity (i.e., congruence) of the factor loadings between groups; (c) if the factors are similar, then conduct the EFA in the combined sample; (d) correct each test's *g* loading for unreliability; (e) standardize the differences in mean scores between the groups; (f) correct each standardized group difference for unreliability; and (g) calculate the correlation (either Pearson or Spearman) between the corrected standardized group differences and the corrected *g* loadings (Jensen, 1985, 1992, 1998). A positive correlation indicates that tests with higher *g* loadings have larger group differences in mean test scores. There is no agreed-upon correlation value that differentiates the strong and weak forms of SH, however, hence support for *g*'s role in determining group differences can vary greatly between studies (Dolan, Roorda, & Wicherts, 2004).

1.4.1.1. Criticisms of the correlated vector method. Scholars have leveled a number of criticisms against the use of a CV analysis to investigate SH (Ashton & Lee, 2005; Mulaik, 1992; Schönemann, 1997). For example, Colom and Lynn (2004) argued that subtest *g* loadings are heavily influenced by the nature of the other subtests included in the battery (see Jensen & Weng, 1994), hence comparing CV studies that have used different instruments to evaluate *g* may be problematic. Dolan and Hamaker (2001) argued that the CV procedure does not adequately assess model fit, thus the factor model used to obtain *g* loadings may not be the best way to explain the tests' covariances. Dolan (2000) opined that making a persuasive argument for *g* as the main contributor to any group differences requires comparing competing models, with the models ascribing a central role to *g* fitting the data better than the models that do not ascribe such a role to *g*.

From a somewhat different perspective, Dolan and colleagues (Dolan, 2000; Dolan & Hamaker, 2001; Dolan & Lubke, 2001; Lubke, Dolan, & Kelderman, 2001) argued that the correlations obtained in a CV analysis are difficult to interpret with any degree of specificity, as the method assumes that the tests are at least strongly invariant across the comparison groups. Strong invariance signifies that any observed group differences in mean test scores are due to group differences in the constructs that the tests are measuring, not differences in how the test measures the construct across groups (i.e., test bias). Thus, if the invariance assumption cannot be established, then between-group differences may be attributable, at least in part, to differences in how the tests measure their intended constructs. Even if invariance holds across groups, when the tests measure multiple factors (e.g., Wechsler scales), CV analysis could mask group differences in lower-order/domain-specific latent variables by implying that the differences are only due to *g*.

1.4.2. Multi-group confirmatory factor analysis method

The multi-group confirmatory factor analysis (MG-CFA) procedure for assessing group differences involves conducting a confirmatory factor analyses (CFAs) simultaneously on separate data from two or more comparison groups (Harrington, 2009). MG-CFA is a well established method for investigating group differences in the latent means and (co)variances estimated from a latent variable model (Millsap, 2011). Moreover, MG-CFA has a number of advantages over a CV analysis for testing SH (Dolan, 2000; Gustafsson, 1992; Horn, 1997; Millsap, 1997).

First, MG-CFA allows for a more integrated and elegant investigation of the various steps involved in the CV analysis. Specifically, MG-CFA requires fitting a single latent variable model in all groups *simultaneously* using the group-specific data. Then, in a systematic fashion the model parameters are constrained to be the same across groups, starting with the factor structure (*configural invariance*), then the loadings (*weak invariance*), and then the intercepts (*strong invariance*). Some (e.g., Lubke, Dolan, Kelderman, & Mellenbergh, 2003) have advocated a need for assessing the equality of the residual variances, too (*strict invariance*), but there is no universal agreement on this (Little, Card, Slegers, & Ledford, 2007). If the loadings and intercepts (i.e., the predicted mean of the observed test for a given level of the latent variable) are the same across groups, then the between-group differences on the measured test scores are only due to between-group differences in the latent means, as opposed to measurement bias playing a role in the observed differences. If the residual

variances are invariant as well, then the reliability with which the test scores measure the latent variables is the same across groups (Raykov, 2004).

Second, the hypothesis of strong factorial invariance—a necessity for meaningful interpretation of group differences—is tested explicitly in MG-CFA. The CV method assumes strong invariance, but only assesses for weak invariance via loading congruence; moreover, this assessment is done in an ad hoc fashion.

Third, MG-CFA can compare models that have different constraints on the model parameter between groups and then use measures to compare how the models fit the data. In the context of SH, this can be advantageous for testing models that include *g* in a central role in explaining group differences against competing models in which *g* does not play a central role in explaining group differences.

1.5. Using multi-group confirmatory factor analysis to examine Spearman's hypothesis

Some investigations of SH have used MG-CFA. Dolan (2000) applied MG-CFA to standardization data for the Wechsler Intelligence Scale for Children-Revised (WISC-R), which Jensen and Reynolds (1982) previously analyzed using the CV approach. Dolan found support for strict invariance between Black and White groups, lending support to the notion that the WISC-R's subtest scores reflected unbiased measurement. They were equivocal about the prominence of *g* causing the group differences, however, because the first- and higher-order factor models that they used to represent the different forms of SH fit the data similarly.

Dolan and Hamaker (2001) used MG-CFA to re-analyze Naglieri and Jensen's (1987) WISC-R and Kaufman Assessment Battery for Children (K-ABC) data. Like Dolan (2000), they found support for strict factorial invariance between Black and White groups. Also like Dolan, they fit multiple first- and higher-order factor models to represent the different forms of SH and could not determine what one fit the data best. Thus, they were equivocal about *g*'s influence on the observed group differences in the test scores. Although Naglieri and Jensen found a CV-based correlation of .75 between *g* and the magnitude of Black–White differences, Dolan and Hamaker concluded that the "repeated demonstration of a positive and large Spearman correlation is a necessary, but not a sufficient condition for inferring the correctness of Spearman's hypothesis" (p. 33).

Not all MG-CFA studies of SH have found support for invariance. For example, Dolan et al. (2004) reanalyzed data from two SH studies (Lynn & Owen, 1994; te Nijenhuis & van der Flier, 1997) that used the CV approach. For both datasets, Dolan et al. did not find evidence for strong invariance and concluded that no form of SH could be inferred from either dataset.

Despite the advantages of the MG-CFA method, this method also has critics. For example, Woodley, te Nijenhuis, Must, and Must (2014) argued that MG-CFA requires large datasets, so studies of SH that used small datasets "simply cannot be analyzed, hence the information contained in them is lost for the purposes of accumulation" (p. 30). Second, MG-CFA cannot be used for a meta-analysis of SH because most studies do not report sufficient information (i.e., within-group means, correlations, and standard deviations). Consequently, they argue that the CV approach is better for examining SH—at least when meta-analytically combing data from multiple studies.

1.6. Factor models used to test Spearman's hypothesis

Studies that have examined SH fall into two groups: those that use a MG-CFA approach and those that use a CV approach. The MG-CFA studies all used a higher-order factor model to represent *g*. Studies that used CV measured *g* in a variety of ways, ranging from the first component of a principal components analysis, to the first unrotated factor from an EFA, to the general factor extracted from Schmid and Leiman's (1957) orthogonal transformation. We contend that none of these are the optimal way to model *g* for an investigation of SH.

1.6.1. Higher-order factor models

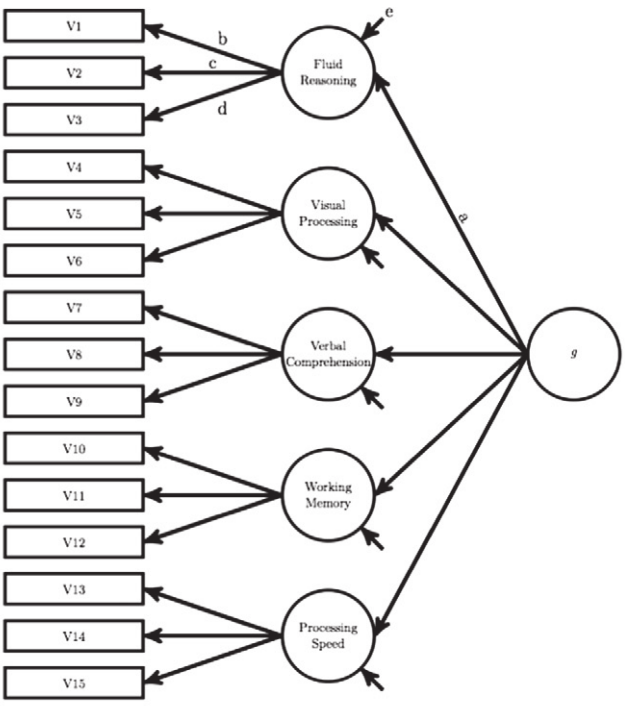
To explain factor models, we use Carroll's (1993, 1996) strata terminology and conceptualization. At Stratum I are narrow factors, which influence a homogenous group of intellectual tasks. There are many factors at Stratum I, some examples of which are Inductive Reasoning, Lexical Knowledge, and Working Memory. At Stratum II are approximately 10 broad factors, which influence a wider range of intellectual tasks than Stratum I factors. Some examples of Stratum II factors are Fluid Reasoning and Comprehension Knowledge. At Stratum III is the single *g* factor, which influences a greater range and diversity of intellectual tasks than any other factor.

The difference between the strata is breadth of content. This is because the presence of factors at a given strata depends on the data being analyzed. If the variables are sufficiently diverse, then *g* will likely be present; with datasets containing variables with homogenous content (e.g., alternate forms of a single test), typically only Stratum I factors are present. For the current study, we only focus on Stratum II and Stratum III because factors derived from individually-administered tests of cognitive ability can typically be classified at one of those strata (Carroll, 1995).

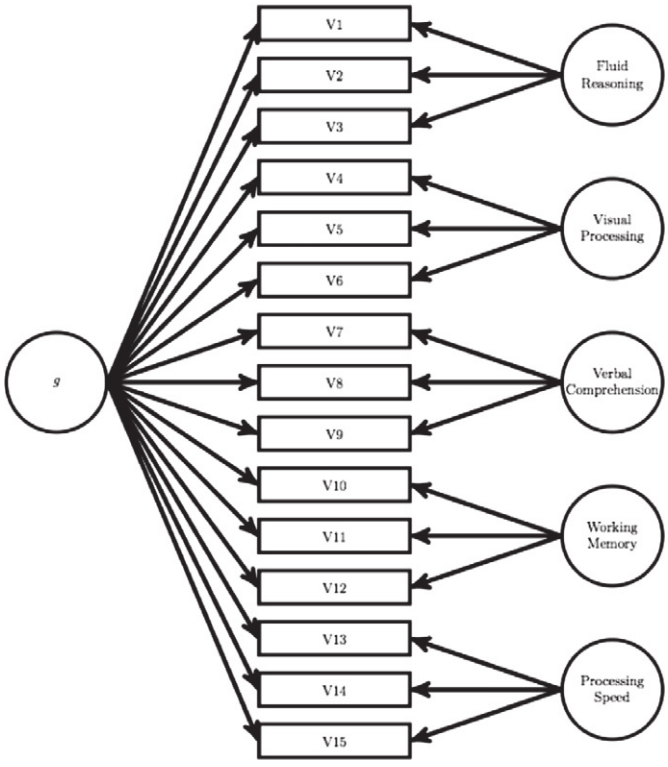
To date, the studies that have examined SH using the MG-CFA approach have all used a higher-order factor (HOF) model (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001). HOF models of cognitive ability define *g* as a single Stratum III (second-order) factor that explains all the common variance among the Stratum II (first-order) factors (see Fig. 1a). The observed test scores have three direct influences: Stratum II factors, test-specific factors, and measurement error. The test-specific factors typically cannot be distinguished from measurement error, so they are amalgamated into a single residual term that is uncorrelated with all other factors.

In HOF models, *g* directly influences all the Stratum II factors. To the extent that *g* is highly correlated with a Stratum II factor, higher levels of *g* produce higher levels of the Stratum II factor. *g* does not directly influence the observed test scores. Instead, *g*'s influence on the tests is mediated by the Stratum II factors.

Stratum II factors can be decomposed into two components in HOF models: the part due to *g* and the part independent of *g*. The part that is independent of *g* is the Stratum II-specific factor, which explains individual differences in the ability that the Stratum II factor represents beyond what *g* can explain. Like the test-specific factors, the Stratum II-specific factors are



(a) Higher-order factor model. *a* is a second-order (Stratum III) factor loading; *b*, *c*, and *d* are first-order (Stratum II) factor loadings; and *e* is Fluid Reasoning's Stratum II-specific variance.



(b) Bi-factor model.

Fig. 1. Intelligence factor models. Test-specific/error variances are not shown for space considerations. While the meaning of the Stratum II factors changes from Model 1a to Model 1b, we have kept the names the same to aid in comparing the two models.

residuals and are typically uncorrelated with all other variables. The *total variance* of a Stratum II factor, then, is an amalgam of the variance attributable to *g* and that attributable to Stratum II-specific factors.

1.6.1.1. Problematic issues associated with higher-order factor models. There are multiple drawbacks of HOF factor models when studying a multidimensional trait such as intelligence (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Gignac, 2007). In these models, *g* does not directly influence the test scores. Thus, its influence on the test scores is limited by how well the test scores measure the Stratum II factors. Fig. 1a illustrates this principle. The relationship between *g* and V1 can be calculated using the tracing rules for a path model (Loehlin, 2004). Specifically, the relationship is calculated by multiplying V1's loading on Fluid Reasoning by Fluid Reasoning's loading on *g* (i.e., $b \times a$). If $b = .30$ and $a = .50$, then the magnitude of *g*'s relationship to V1 is $.30 \times .50 = .15$. If *b* increases to .65, then *g*'s relationship with V1 increases to $.65 \times .50 = .33$.

Another drawback of HOF models is that they impose proportionality constraints (Yung, Thissen, & McLeod, 1999). Specifically, for a given set of tests influenced by the same Stratum II factor, the ratio of the test scores' variance due to the Stratum II factor to the variance attributable to *g* is constrained to be the same.

Proportionality constraints can be a challenge to understand clearly, so we follow Beaujean, Parkin, and Parker's (2014) explanation using Fig. 1a. We previously showed how to calculate the relationship between *g* and V1 using tracing rules. We can use the same tracing rules to compute the influence of Fluid Reasoning's Stratum II-specific factor on V1. Specifically, multiply V1's loading on Fluid Reasoning by the standard deviation of Fluid Reasoning's Stratum II-specific factor (e.g., $b \times \sqrt{e}$). The ratio of *g*'s indirect influence on V1 to the influence of Fluid Reasoning's Stratum II-specific factor on V1 is exactly the same for the other observed test scores that Fluid Reasoning influences: V2, and V3. Specifically,

$$\frac{b \times a}{b \times \sqrt{e}} = \frac{c \times a}{c \times \sqrt{e}} = \frac{d \times a}{d \times \sqrt{e}}$$

These forced proportional loading patterns can be problematic. First, the constraints cause multicollinearity problems when using both *g* and Stratum II factors as predictor variables (Beaujean et al., 2014). Second, it is unlikely that such constraints occur in a population (Schmiedek & Li, 2004). Although some have empirically assessed the tenability of proportionality constraints and not found them problematic (e.g., Dolan & Hamaker, 2001), Mulaik and Quartetti (1997) argued that the sample sizes needed for such investigations are much larger than what is typically used in SH investigations. Third, proportionality constraints confound *g* and Stratum II factors in HOF models because the second-order factor structure is just a re-expression of the Stratum II factors correlations (Reise, 2012). A combination of the last two issues could possibly explain why previous SH studies found equivalent fit between HOF models and oblique first-order models, and, subsequently, could not determine if group differences were due to *g* or Stratum II factors.

These criticisms apply just as well to any transformation of HOF model such as the one developed by Schmid and Leiman (1957). While the Schmid–Leiman transformation can aid in the interpretation of higher-order model's Stratum II factors, it does not release the proportionality constraints. It is only through a bi-factor model that the Stratum II factors' constraints on *g* are released.

1.6.2. Schmid–Leiman transformation (SLT)

Schmid and Leiman (1957) developed a matrix transformation that some use with higher-order models to calculate all the direct and indirect influences on the indicator variables simultaneously (Reynolds & Keith, 2013). Another use of the Schmid–Leiman transformation (SLT) is to combine the results from an EFA on observed test scores (i.e., first-order EFA) that have oblique (correlated) factors and an EFA of the correlated factors (i.e., second-order EFA; Gorsuch, 1983). In either case, the SLT produces *g* loadings for the observed test scores via the technique discussed in Section 1.6.1.1.

In the SLT, the common variance among all the test scores is represented as a general factor, while narrower domains are represented as residual Stratum II factors. Consequently, the Stratum II factors are orthogonal to each other as well as to the general factor. Thus, Stratum II factors from a SLT do not have the same interpretation as those from a Stratum II EFA with an oblique rotation. In the oblique rotation, the Stratum II factors reflect variance from both *g* and the Stratum II factors, whereas in the SLT the Stratum II factors only reflect variance at the Stratum II factor level that is unexplained by *g* (Reise, 2012). Despite the differences in factor construction, the convention has been to call Stratum II factors by the same name regardless of how they were formed (e.g., Carroll, 1996).

1.6.2.1. Problems with the Schmid–Leiman transformation. There are two major problems with the SLT. First, the direct factor loadings produced by the SLT are merely a re-expression of the correlations among the Stratum II factors. Thus, the factor loadings of an EFA with correlated Stratum II factors and a SLT of the EFA's factor loading are equivalent (Schmid, 1957); the same can also be said for the loadings from a higher-order CFA and a SLT of those loadings (Yung et al., 1999). Consequently, the SLT does not do away with the proportionality constraints in a HOF model and imposes the constraints on the second-order EFA.

A second major problem of the SLT occurs when there are cross-loadings (i.e., some of the observed tests load onto more than one Stratum II factor), which are not uncommon with individually-administered intelligence tests (Weiss, Keith, Zhu, & Chen, 2013a, 2013b). In such situations, the SLT will overestimate the *g* loadings and underestimate the Stratum II factor loadings (Reise, Moore, & Haviland, 2010). Larger Stratum II cross-loadings produce larger amounts of the over- or underestimation (Reise, 2012).

1.6.3. Bi-factor models

The bi-factor model (Holzinger & Swineford, 1937), sometimes called a direct hierarchical or nested-factors model, offers an alternative to both the HOF in the MG-CFA approach and the second-order EFA in the CV approach.[2]

[2] Technically, the bi-factor model is a generalization of the HOF model (Gignac, 2008; Yung et al., 1999), but we consider them as two distinct models.

1.6.3.1. Bi-factor model for confirmatory factor analysis. An example of a bi-factor (BF) model is shown in Fig. 1b. In this model, *all* factors have a direct influence on the tests. Consequently, higher levels of *g* and higher levels of the Stratum II factors are both directly associated with higher scores on all the tests (assuming positive loadings). The difference between *g* and Stratum II factors is that while *g* is thought to influence every test, the Stratum II factors only influence a subset of the tests. Test-specific factors and measurement error also influence the tests in the BF model. As with the HOF model, the test-specific factors' influence is typically indistinguishable from influence due to measurement error so they are represented as a single residual term that is uncorrelated with any other factor.

BF models have advantages over the HOF (Chen, West, & Sousa, 2006). First, unlike the HOF, the BF model forms the Stratum II factors from the covariance remaining after accounting for *g*, making the Stratum II factors independent of *g* (i.e., are all uncorrelated). Thus, the BF model produces a direct estimation of the relationship between the observed tests scores and Stratum II-specific factors. Second, the BF model allows the tests' factor loadings on both *g* and the Stratum II factors to be estimated without any proportionality constraints.

A third advantage of the BF model is that is allows for direct assessment of measurement invariance in both *g* and the Stratum II factors. In HOF models, non-invariance of a Stratum II factor would automatically produce non-invariance in *g*. Fourth, the BG model allows for a direct comparison of mean differences between groups on Stratum II factors independent of *g*. These last two advantages are particularly salient when examining SH. If there is at least strong invariance in *g* and the Stratum II factors, then the BF model allows for a simultaneous investigation of the strong, weak, and contra forms of SH.

Specifically, support for the strong form of SH would come from there being no differences in the latent mean of the Stratum II factors, but there being a difference in the latent mean of *g*. Conversely, support for the contra hypothesis would come from there being differences in the latent mean of the Stratum II factors but no difference in the latent mean of *g*. If there were differences in the latent means of both *g* and the Stratum II factors, then this would provide evidence of the weak form of SH.

1.6.3.2. Bi-factor rotations for exploratory factor analysis. Recently, two bi-factor methods have been developed for EFA. The first is bi-factor target rotations (Reise, Moore, & Maydeu-Olivares, 2011). The basic idea is to extract factors as usual in an EFA, specify a factor pattern matrix to use for factor rotation, and then rotate the factors to minimize the difference between the estimated factor loadings and the specified elements of the target factor loadings. For more information on target rotation, see Browne (2001).

The second BF method for EFA is an analytic rotation (Jennrich & Bentler, 2011, 2012). Here EFA is done as usual, only the factors are rotated such that all the tests load on the first factor and the remaining factors are rotated in such a way to encourage perfect cluster structure (i.e., the tests have substantial loadings on only one factor). The first factor is the general factor and is uncorrelated with the other factors. The remaining factors can either be correlated or uncorrelated with each other.

1.7. Purpose of the current study

The purpose of the current investigation is to test SH using BF models and both CV and MG-CFA approaches. To do this, we used Black and White adults' scores from the co-normed Wechsler Adult Intelligence Scale-Fourth Edition (Wechsler, 2008a) and Wechsler Memory Scale-Fourth Edition (Wechsler, 2009) standardization data.

Based on our review of the SH literature, we expect to find support for the weak form of SH in this dataset. The reason is that the dataset contains tests that have both high and low *g* loadings and oversamples tests of memory. Thus, group differences in the test scores are likely due to group differences in both *g* and Stratum II factors. If our hypothesis is correct, the CV analysis will produce a moderately sized positive correlation between the tests' *g* loadings and the size of Black–White test score differences. In the MG-CFA analysis, support for the weak form of SH would come from mean differences in *g* favoring the White sample, but small or no group differences in non-memory Stratum II factors. Mean differences on any memory factors should either show no Black–White difference or, if a difference exists, favoring the Black sample (as suggested from previous research).

2. Method

2.1. Materials

2.1.1. Wechsler Adult Intelligence Scale-Fourth Edition

The Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008a) is an individually administered battery designed to assess cognitive ability in individuals between the ages of 16–90 years. The WAIS-IV consists of 10 primary subtests (Vocabulary, Information, Similarities, Digit Span, Arithmetic, Block Design, Matrix Reasoning, Visual Puzzles, Coding, and Symbol Search). The primary subtests yield four Index scores (Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed) and an overall Full-Scale IQ. The average internal consistency reliability of WAIS-IV subtests ranged from .78 for to .94 (Wechsler, 2008b).

2.1.2. Wechsler Memory Scale-Fourth Edition

The Wechsler Memory Scale-Fourth Edition (WMS-IV; Wechsler, 2009) is an individually administered battery designed to assess a variety of memory abilities, such as working memory, learning, immediate and delayed recall, and recognition of information. There are both verbal and visual tasks are presented in verbal and visual modalities, and was standardized on individuals between the ages of 16–90 years. Not counting the Brief Cognitive Status Exam, the subtests include Logical Memory (recall for a short story); Verbal Paired Associates (recall for related and unrelated word pairs); Designs (recall of spatial locations and visual details); Visual Reproduction (recall of geometric designs); Spatial Addition (ability to manipulate visual–spatial information in working memory); and Symbol Span (ability to manipulate designs in working memory). The average internal consistency reliability

of these subtests ranged from .82 to .97 (Wechsler, Holdnack, & Drozdick, 2009).

2.2. Participants

Participants were members the WAIS-IV and WMS-IV co-normative sample, which is made up of adults aged 16 through 90 years. The sample closely matched the 2005 census on gender, age, race/ethnicity, parent education level, and geographic region. For more information about the sample, see Wechsler et al. (2009). There were 1250 total respondents, 1015 of whom identified as either Black ($n = 180$) or White ($n = 835$). Only the Black and White respondents were used for this study. Descriptive statistics for the subtest scores are given in Table 1.

2.2.1. Missing data

There were 737 respondents with no missing data on any of the WAIS-IV subtests, 1 respondent missing a score on the Picture Completion subtest, 1 respondent missing a score on the Cancellation subtest, and 276 respondents missing data on the Figure Weights, Letter-Number Sequencing, and Cancellation subtests, almost all of whom were age 70 or above. There were 700 respondents with no missing data on any of the WMS-IV subtests and 315 missing data on the Designs and Spatial Addition subtests, all age 70 or above.

There were 699 respondents with no missing data on the WAIS-IV or WMS-IV subtests, 1 respondent missing only the score on the Cancellation subtest, 38 respondents missing data on only the Designs and Spatial Addition subtests, and 1 respondents missing data on the Picture Completion, Designs, and Spatial Addition subtests. In addition, there were 276 respondents missing data on Figure Weights, Letter-Number Sequencing, Cancellation, Designs, and Spatial Addition subtests, all age 70 or above.

The majority of the missing data are missing due to the design of the data collection (e.g., planned missingness; McArdle, 1994). That is, respondents above the age of 70 years were not administered the Figure Weights, Letter-Number Sequencing, Cancellation, Designs, and Spatial Addition subtests. While deleting respondents with missing values on these variables (i.e., only keeping respondents younger than 70 years) would likely not bias the results, we instead chose to handle the missing data using full information maximum likelihood (FIML) estimation (Enders & Bandalos, 2001). Unlike traditional ML estimation, FIML makes use of all the data available from each respondent so respondents do not have to be removed from the dataset because they were missing values.

2.3. Data analysis

We tested SH using two methods, the CV approach (Method 1) and a MG-CFA approach (Method 2).

2.3.1. Method 1: correlated vector analysis

We followed the steps outlined in Section 1.4.1, with some modifications. First, we group centered the variables (i.e., created mean deviation scores separately for the Black and White groups) before conducting the EFA in the combined group. Second, as there were missing values in the data, we first created FIML-based correlation matrices of all the WAIS-IV and WMS-IV subtests for the Black, White, and combined groups.

Table 1
Descriptive statistics for Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) and Wechsler Memory Scale-Fourth Edition (WMS-IV) subtests.

Battery	Subtest	*n*	Mean	SD	Skew	Kurtosis
WAIS-IV	Block Design	1015	10.22	2.98	0.02	−0.35
WAIS-IV	Matrix Reasoning	1015	10.30	3.02	−0.01	−0.50
WAIS-IV	Figure Weights	739	10.40	3.02	0.01	−0.30
WAIS-IV	Picture Completion	1014	10.14	3.03	−0.22	−0.39
WAIS-IV	Symbol Search	1015	10.14	2.91	0.09	0.05
WAIS-IV	Coding	1015	10.17	2.92	−0.03	−0.12
WAIS-IV	Cancellation	738	9.99	2.89	0.38	0.20
WAIS-IV	Vocabulary	1015	10.30	2.92	0.13	−0.14
WAIS-IV	Information	1015	10.06	3.05	0.06	−0.50
WAIS-IV	Comprehension	1015	10.49	3.07	−0.04	−0.22
WAIS-IV	Similarities	1015	10.28	2.85	−0.15	0.01
WAIS-IV	Arithmetic	1015	10.24	2.85	0.10	−0.42
WAIS-IV	Digit Span	1015	10.26	2.80	0.19	−0.02
WAIS-IV	Letter-Number Sequencing	739	10.39	3.00	0.82	1.47
WAIS-IV	Visual Puzzles	1015	10.19	3.05	0.38	−0.46
WMS-IV	Logical Memory I	1015	10.10	3.02	−0.26	−0.24
WMS-IV	Logical Memory II	1015	9.95	2.94	−0.18	0.01
WMS-IV	Visual Reproduction I	1015	10.12	3.09	−0.32	0.12
WMS-IV	Visual Reproduction II	1015	10.02	3.12	0.13	0.24
WMS-IV	Verbal Paired Associates I	1015	9.91	2.98	0.06	−0.24
WMS-IV	Verbal Paired Associates II	1015	9.88	3.00	−0.41	−0.19
WMS-IV	Designs I	700	10.11	3.01	−0.06	−0.24
WMS-IV	Designs II	700	10.01	3.04	−0.07	0.06
WMS-IV	Spatial Addition	700	10.07	3.07	−0.26	−0.43
WMS-IV	Symbol Span	1015	10.07	3.00	−0.03	−0.35

We used these correlation matrices for the EFAs. To calculate *g*-loadings, we used the analytic bi-factor rotation. To assess factor similarity between the Black and White groups, we estimated the congruence coefficients (Lorenzo-Seva & ten Berge, 2006), with values ≥ .95 indicating sufficient similarity for *g* and values ≥ .85 indicating sufficient similarity for the other factors (te Nijenhuis & van der Flier, 2003).

We measured the Black–White standardized difference by calculating Hedges (1981) effect size (ES) measure, which expresses the mean difference between groups in standard deviation units. We used Hedges' ES as it corrects for the slight bias in the more commonly used *d* effect measure (Borenstein, 2009). The ES formula is given in Eq. (1).

$$\mathrm{ES} = \left(1 - \frac{3}{4(df)-1}\right) \frac{\bar{x}_W - \bar{x}_B}{\sqrt{\frac{(n_w-1)s_W^2 + (n_B-1)s_B^2}{df}}} \quad (1)$$

where, $\bar{x}_B$ and $\bar{x}_W$ are the mean scaled scores of the Black and White groups, respectively, s_B^2 and s_W^2 are the respective variances for the Black and White groups, and *df* are the degrees of freedom calculated as $n_W + n_B - 2$, where n_i is the *i*th group's sample size.

We conducted the CV analyses using two versions of the WAIS-IV/WMS-IV data. In the first version, we used all subtests within both the WAIS-IV and WMS-IV batteries. Due to statistical problems experienced by previous researchers using both immediate and delayed versions of the WMS-IV subtests (see Section 2.3.2), in the second analysis we omitted the WMS-IV immediate subtests. This second analysis was conducted in order to compare the results from the CV and MG-CFA analyses.

2.3.2. Method 2: multi-group confirmatory factor analysis

Independent factor analytic studies of the WAIS-IV (e.g., Benson, Beaujean, & Taub, in press; Benson, Hulac, & Kranzler, 2010; Gignac & Watkins, 2013; Nelson, Canivez, & Watkins, 2013; Niileksela, Reynolds, & Kaufman, 2013; Ward, Bergman, & Hebert, 2012; Wechsler, 2008b) have shown the scale to reflect four or five latent variables, mapping onto either the four WAIS-IV index scores (Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed) or the Cattell–Horn–Carroll (Schneider & McGrew, 2012) theory (Comprehension Knowledge, Visual Processing, Fluid Reasoning, Short Term Memory, and Processing Speed) respectively. The difference in factor models between studies likely comes from whether the model allowed the subtests to have cross-loadings. Weiss et al. (2013a) argued that the four- and five-factor models were both sufficient for demonstrating model fit and full factorial invariance between clinical and nonclinical samples.

There have been some difficulties in forming CFA models with WMS data (Wechsler et al., 2009, p. 6). In the WMS-IV, some subtests require examinees to recall stimuli immediately after presentation (subtests comprising the Immediate Index), while other subtests ask examinees to recall this stimuli after a delayed period of time after which intervening and nonrelated subtests have been administered previously (subtests comprising the Delayed Index). The difficulty including both Immediate and Delayed WMS-IV subtests is that it produces specification errors and inadmissible parameter estimates (Millis, Malina, Bowers, & Ricker, 1999; Price, Tulsky, Millis, & Weiss, 2002). Thus, most factor analyses of the WMS-IV data (e.g., Holdnack, Zhou, Larrabee, Millis, & Salthouse, 2011; Miller, Davidson, Schindler, & Messier, 2013; Salthouse, 2009), including those in the WMS-IV technical and interpretive manual (Wechsler et al., 2009, p. 60), only include one version of these subtests (usually the delayed) along with the two visual working memory measures (Spatial Addition and Spatial Span).

There have been a few studies examining the factor structure of the WAIS-IV and WMS-IV subtests concurrently. Holdnack et al. (2011) completed the most thorough study, examining thirteen different models in the 900 participants from the co-norming sample between the ages of 16–69 years. They found two models fit the data relatively well. The first model included seven Stratum II factors (Verbal Comprehension, Perceptual Reasoning, Processing Speed, Auditory Working Memory, Visual Working Memory, Auditory Memory, and Visual Memory) without a *g* factor. The second model contained five Stratum II factors (Verbal Comprehension, Perceptual Reasoning, Processing Speed, Working Memory, and Long-Term Retrieval) and a higher-order *g* factor.

Miller et al. (2013) analyzed WAIS-IV/WMS-IV data and found a model similar to Holdnack et al.'s (2011) five-factor model. Specifically, they found five Stratum II factors (Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed, and Delayed Memory) and a higher-order *g* factor fit their data best. Salthouse's (2009) found that a model with six Stratum II factors and a higher-order *g* factor fit the WAIS-IV/WMS-IV analysis best. Four factors were the same as those from Miller et al.'s and Holdnack et al.'s studies (Verbal Comprehension, Fluid Reasoning, Working Memory, Processing Speed). The difference is that Salthouse's model splits the Delayed Memory/Long-Term Retrieval factor into two separate factors: Verbal Memory, and Visual Memory. Likely, this difference stems from Salthouse using the immediate version of the WMS-IV subtests instead of the delayed.

For our MG-CFA study, we used all WAIS-IV subtests and only the delayed and visual working memory subtests from the WMS-IV. We chose the delayed subtests over the immediate tests because those are the ones most commonly used and are the ones used in the factor analyses reported in the WMS-IV technical and interpretive manual (Wechsler et al., 2009). Our investigation differs from previous investigations in that we used a BF model to extract *g* and tested for invariance between the Black and White groups before evaluating SH.

2.3.2.1. Determining model fit. To determine model fit, we examined multiple indices (McDonald & Ho, 2002) that represent a variety of fit criteria (Marsh, Hau, & Grayson, 2005). Specifically, we examined (a) the χ^2, (b) the comparative fit index (CFI), (c) root mean square error of approximation (RMSEA), (d) standardized root mean square residual (SRMR), and (e) Mcdonald's non-centrality index (Mc). In addition, following Boomsma's (2000) recommendation we also reported Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). For all models, we looked for patterns in the fit statistics, and judged acceptance/rejection of

the specific model based on the majority of the indices. We used the following criteria for overall model fit: (a) CFI ≥.96 (Yu, 2002); (b) RMSEA ≤.08 (Browne & Cudeck, 1993; Hu & Bentler, 1999); (c) SRMR ≤.08, (Hu & Bentler, 1999; Sivo, Fan, Witta, & Willse, 2006); (d) Mc > .90. While AIC and BIC measure different aspects of model fit, for both measures smaller values indicate better approximations of the true model (Markon & Krueger, 2004). When specifically comparing models for invariance, Meade, Johnson, and Braddy (2008) suggest that changes in CFI values of .002 and changes in Mc values between .008–.009 are useful cutoff points.

All analyses were conducted using the R statistical programming language (R. Development Core Team, 2014), using the *psych* (Revelle, 2012) and *lavaan* (Rosseel, 2012) packages. To conduct the EFA, fit the latent variable models, and assess invariance in R, we followed the steps outlined in Beaujean (2014a, 2014b).

3. Results

3.1. Method 1: correlated vectors

3.1.1. All subtests

We conducted the first step of the CV analyses using all subtests in the WAIS-IV and WMS-IV co-normed dataset. Velicer's (1976) minimum average partial (MAP) criterion suggested 2 factors, while Horn's (1965) parallel analysis suggested 4–8 factors. We believe that the seven-factor solution made the most interpretive sense (see Table 2). The extracted factors consist of *g*, a Verbal Comprehension factor (F1), a Logical Memory factor (F2), a Designs factor (F3), a Verbal Paired Associates factor (F4), a Processing Speed factor (F5), and a Short–Term Memory factor (F6).

We calculated the congruence coefficient (CC) from EFAs applied to Black and White groups separately. The CC was 1.00 for *g* and had values ≥.87 for all the other factors. Since the CCs were sufficiently high, we combined Black and White samples and extracted the *g* loadings from an analytic BF rotation. The subtest data we used for the correlated vectors (CV) analysis are given in Table 3. The Pearson correlation between the corrected *g* loadings and the corrected Black-White standardized differences was 0.58 (95% CI: 0.53–0.62), while the Spearman correlation was .62 (95% CI: .58–.66).

3.1.2. All subtests except WMS-IV immediate

We performed this analysis to mirror the multi-group CFA we conducted. MAP criterion suggested 2 factors, while parallel analysis suggested 2–5 factors. We believe that the five-factor solution made the most interpretive sense (see Table 4). It is comprised of *g*, a Comprehension Knowledge factor (F1), a Processing Speed factor (F2), a Working Memory factor (F3), and a Visual Spatial Factor (F4).

The CC for *g* was 1.00, and was ≥.89 for all the others factors. Since the CCs were sufficiently high, we combined Black and White samples and extracted the *g* loadings from an analytic BF rotation. The subtest data we used for the CV analysis are given in Table 5. The Pearson correlation between the corrected *g* loadings and the corrected standardized Black–White differences is 0.57 (95% CI: 0.53–0.61), while the Spearman correlation is .65 (95% CI: .61–0.68).

3.2. Method 2: multi-group confirmatory factor analysis

3.2.1. Testing assumptions

A major assumption of SEM is that the manifest variables are multivariate normal (Kline, 2012). All WAIS-IV and WMS-

Table 2
Results from exploratory factor analysis of all subtests extracting seven factors using the combined sample ($n = 1015$).

		Factor pattern coefficients						
Battery	Test	*g*	F1	F2	F3	F4	F5	F6
WAIS-IV	Figure Weights	0.71	0.11	−0.12	−0.03	−0.05	−0.02	0.05
WMS-IV	Visual Reproduction I	0.68	−0.06	−0.11	0.11	−0.02	0.04	−0.21
WAIS-IV	Visual Puzzles	0.67	−0.06	−0.20	0.00	−0.15	−0.02	−0.14
WAIS-IV	Matrix Reasoning	0.66	0.04	−0.15	0.04	−0.07	0.04	0.00
WMS-IV	Symbol Span	0.66	0.00	−0.03	0.18	0.02	0.05	0.03
WAIS-IV	Block Design	0.65	−0.02	−0.21	0.04	−0.18	0.02	−0.13
WAIS-IV	Arithmetic	0.65	0.19	−0.04	−0.04	−0.07	−0.01	0.21
WAIS-IV	Digit Span	0.62	0.05	−0.04	0.00	−0.04	0.08	0.55
WAIS-IV	Similarities	0.61	0.48	0.03	−0.08	−0.04	−0.06	0.01
WMS-IV	Spatial Addition	0.61	0.02	−0.14	0.12	−0.08	0.10	0.07
WAIS-IV	Vocabulary	0.59	0.66	0.05	−0.03	0.04	0.00	0.05
WMS-IV	Visual Reproduction II	0.59	−0.12	−0.05	0.15	0.03	0.00	−0.20
WAIS-IV	Comprehension	0.59	0.52	0.05	−0.02	−0.03	−0.03	0.02
WAIS-IV	Information	0.59	0.50	−0.02	−0.06	−0.07	−0.04	−0.05
WMS-IV	Verbal Paired Associates I	0.58	−0.02	0.08	0.04	0.68	−0.01	0.01
WMS-IV	Logical Memory I	0.58	0.03	0.69	−0.03	0.03	−0.03	0.02
WAIS-IV	Letter–Number Sequencing	0.58	0.03	−0.09	−0.02	−0.03	0.03	0.46
WAIS-IV	Picture Completion	0.58	0.00	−0.01	0.00	−0.09	0.09	−0.08
WMS-IV	Verbal Paired Associates II	0.55	−0.01	0.07	0.03	0.74	−0.03	−0.03
WMS-IV	Logical Memory II	0.54	0.02	0.76	−0.03	0.09	−0.01	−0.05
WAIS-IV	Coding	0.52	0.00	0.00	0.02	−0.03	0.57	0.08
WAIS-IV	Symbol Search	0.51	−0.04	−0.04	0.01	−0.04	0.65	−0.01
WMS-IV	Designs I	0.50	−0.04	−0.03	0.74	0.01	0.05	−0.03
WMS-IV	Designs II	0.44	−0.04	−0.02	0.69	0.07	−0.03	0.03
WAIS-IV	Cancellation	0.38	−0.10	−0.04	0.11	−0.03	0.37	0.07

Note. Factors were rotated using analytic bi-factor rotation. Subtests are presented in descending order of their *g* loadings.

Table 3
Data used in correlated vectors analysis of all subtests.

Battery	Test	ES	n_W	n_B	Corrected ES	*g* Loading	Corrected *g* Loading
WAIS-IV	Figure Weights	0.81	590	149	0.86	0.71	0.75
WMS-IV	Visual Reproduction I	0.66	835	180	0.69	0.68	0.71
WAIS-IV	Visual Puzzles	0.85	835	180	0.91	0.67	0.71
WAIS-IV	Matrix Reasoning	0.79	835	180	0.84	0.66	0.70
WMS-IV	Symbol Span	0.62	835	180	0.66	0.66	0.70
WAIS-IV	Block Design	1.19	835	180	1.27	0.65	0.70
WAIS-IV	Arithmetic	0.74	835	180	0.79	0.65	0.69
WAIS-IV	Digit Span	0.62	835	180	0.64	0.62	0.64
WAIS-IV	Similarities	0.80	835	180	0.85	0.61	0.66
WMS-IV	Spatial Addition	0.78	560	140	0.81	0.61	0.64
WAIS-IV	Comprehension	0.84	835	180	0.90	0.59	0.63
WAIS-IV	Vocabulary	0.80	835	180	0.82	0.59	0.61
WAIS-IV	Information	0.78	835	180	0.81	0.59	0.61
WMS-IV	Visual Reproduction II	0.49	835	180	0.49	0.59	0.60
WMS-IV	Logical Memory I	0.62	835	180	0.68	0.58	0.64
WAIS-IV	Picture Completion	0.95	835	179	1.03	0.58	0.63
WAIS-IV	Letter–Number Sequencing	0.60	590	149	0.64	0.58	0.61
WMS-IV	Verbal Paired Associates I	0.47	835	180	0.49	0.58	0.60
WMS-IV	Verbal Paired Associates II	0.46	835	180	0.50	0.55	0.60
WMS-IV	Logical Memory II	0.60	835	180	0.65	0.54	0.58
WAIS-IV	Coding	0.74	835	180	0.80	0.52	0.56
WAIS-IV	Symbol Search	0.72	835	180	0.80	0.51	0.56
WMS-IV	Designs I	0.56	560	140	0.61	0.50	0.55
WMS-IV	Designs II	0.43	560	140	0.46	0.44	0.47
WAIS-IV	Cancellation	0.38	589	149	0.43	0.38	0.43

Note. ES: Hedges' effect size. N_W: White sample size. N_B: Black sample size. Corrected: corrected for unreliability. Subtests presented in descending order of their *g* loading. Scores from the Black participants were subtracted from the White participants, so a positive ES indicates that the average score from the White group was higher.

IV univariate subtest skewness values were < 2 and all kurtosis values were < 7, so were in acceptable limits (West, Finch, & Curran, 1995). While Mardia (1980) tests of multivariate skew and kurtosis were larger than expected ($b_{1,p} = 30.81$, $b_{2,p} = 687.85$, $n = 738$, $p = 21$), a Q–Q plot of the multivariate distribution does not look markedly different from data plotted from a known multivariate normal distributions with the same *n*, number of variables, and correlation matrix (Andersen, 2012). Consequently, we believe that the data approximate a multivariate normal distribution.

3.2.1.1. Confirmatory factor analysis. Based on our EFA, we initially fit a bi-factor model with four Stratum II factors (Verbal Comprehension, Processing Speed, Visual Processing, and Working Memory). The values for the fit statistics for this model (B0) are shown in Table 6. The fit measures meet the

Table 4
Results from exploratory factor analysis of all subtests except the WMS-IV immediate subtests, extracting five factors using the combined sample ($n = 1015$).

		Factor pattern coefficients				
Battery	Test	*g*	F1	F2	F3	F4
WAIS-IV	Figure Weights	0.70	0.09	−0.03	0.11	0.15
WMS-IV	Symbol Span	0.68	−0.09	0.02	0.06	−0.05
WAIS-IV	Matrix Reasoning	0.66	0.01	0.04	0.05	0.16
WAIS-IV	Visual Puzzles	0.66	−0.07	−0.01	−0.06	0.32
WAIS-IV	Block Design	0.65	−0.02	0.02	−0.04	0.36
WAIS-IV	Arithmetic	0.63	0.17	−0.02	0.25	0.10
WAIS-IV	Vocabulary	0.63	0.60	0.02	0.02	−0.06
WAIS-IV	Similarities	0.63	0.48	−0.03	0.01	0.02
WMS-IV	Visual Reproduction II	0.61	−0.19	−0.01	−0.12	−0.06
WAIS-IV	Comprehension	0.61	0.50	−0.01	0.00	−0.03
WMS-IV	Spatial Addition	0.61	−0.03	0.08	0.11	0.13
WAIS-IV	Information	0.61	0.49	−0.01	−0.04	0.09
WAIS-IV	Digit Span	0.58	0.00	0.01	0.63	−0.02
WAIS-IV	Picture Completion	0.56	0.03	0.11	−0.03	0.11
WAIS-IV	Letter–Number Sequencing	0.54	−0.02	−0.02	0.50	0.03
WMS-IV	Verbal Paired Associates II	0.54	−0.08	−0.05	−0.02	−0.38
WAIS-IV	Coding	0.51	0.03	0.59	0.04	−0.02
WAIS-IV	Symbol Search	0.51	−0.01	0.64	−0.04	0.02
WMS-IV	Designs II	0.50	−0.23	−0.07	0.03	−0.11
WMS-IV	Logical Memory II	0.48	0.10	0.05	−0.05	−0.36
WAIS-IV	Cancellation	0.39	−0.13	0.34	0.07	−0.01

Note. Factors were rotated using analytic bi-factor rotation. Subtests presented in descending order of their *g* loading.

Table 5
Data used in correlated vectors analysis of all subtests except the WMS-IV immediate subtest scores.

Battery	Test	ES	n_W	n_B	Corrected ES	*g* Loading	Corrected *g* loading
WAIS-IV	Figure Weights	0.81	590	149	0.86	0.70	0.74
WMS-IV	Symbol Span	0.62	835	180	0.66	0.68	0.73
WAIS-IV	Matrix Reasoning	0.79	835	180	0.84	0.66	0.70
WAIS-IV	Visual Puzzles	0.85	835	180	0.91	0.66	0.70
WAIS-IV	Block Design	1.19	835	180	1.27	0.65	0.69
WAIS-IV	Arithmetic	0.74	835	180	0.79	0.63	0.67
WAIS-IV	Similarities	0.80	835	180	0.85	0.63	0.67
WAIS-IV	Vocabulary	0.80	835	180	0.82	0.63	0.65
WAIS-IV	Comprehension	0.84	835	180	0.90	0.61	0.66
WMS-IV	Spatial Addition	0.78	560	140	0.81	0.61	0.64
WAIS-IV	Information	0.78	835	180	0.81	0.61	0.63
WMS-IV	Visual Reproduction II	0.49	835	180	0.49	0.61	0.62
WAIS-IV	Digit Span	0.62	835	180	0.64	0.58	0.60
WAIS-IV	Picture Completion	0.95	835	179	1.03	0.56	0.61
WMS-IV	Verbal Paired Associates II	0.46	835	180	0.50	0.54	0.59
WAIS-IV	Letter–Number Sequencing	0.60	590	149	0.64	0.54	0.58
WAIS-IV	Symbol Search	0.72	835	180	0.80	0.51	0.56
WAIS-IV	Coding	0.74	835	180	0.80	0.51	0.55
WMS-IV	Designs II	0.43	560	140	0.46	0.50	0.55
WMS-IV	Logical Memory II	0.60	835	180	0.65	0.48	0.52
WAIS-IV	Cancellation	0.38	589	149	0.43	0.39	0.44

Note. ES: Hedges' effect size. n_W: White sample size. n_B: Black sample size. Corrected: Corrected for unreliability. Subtests presented in descending order of their *g* loading. Scores from the Black participants were subtracted from the White participants, so a positive ES indicates that the average score from the White group was higher.

RMSEA and SRMR criteria, but do not meet the Mc criterion and are at the threshold of the CFI criterion. Examining the residual correlations and modification indices indicated that we should include a fifth, Long-Term Retrieval factor, making the model similar to that used by Holdnack et al. (2011) and Miller et al. (2013). In addition, we allowed the residuals between the WAIS-IV Figure Weights and Arithmetic subtests and the residuals between the WMS-IV Logical Memory and Verbal Paired Associates subtests to covary. This new model (B1) fit the data better than the model with four factors, so we used it for our baseline model. A path diagram of the model is shown in Fig. 2. In Model B1, not all subtests loaded on Stratum II factors, indicating that *g* explained all the covariance between those subtests and the other subtests in the dataset.

Next, we fit model B1 to the Black and White groups, separately (B1.B and B1.W, respectively). The model fit slightly better in the White group than in the Black group, although the fit is equivalent in most respects. We then assessed for invariance using the steps listed in Section 1.4.2. The constraints involved in the configural model (M1), weak model (M2), and strong invariance model (M3) did not depreciate the model fit. In fact, the AIC and BIC that showed the model with more constraints fit slightly better than the models without them. Thus, it appears that the factors are comparable across groups.

To examine strict invariance, we added constraints in two parts, one for the residual variances and once for the residual covariances. After constraining the residual variances (M4a),

Table 6
Model fit for combined WAIS-IV and WMS-IV multi-group confirmatory factor models.

Model	Description	χ^2	*df*	*p*	CFI	RMSEA	SRMR	Mc	AIC	BIC
B0	Baseline: 4 Stratum II factors, all respondents	552.662	173	.00	.961	.05	.035	.829	90344	90733
B1	Baseline: 5 Stratum II factors, all respondents	362.938	166	.00	.980	.03	.028	.908	90168	90592
B1.B	Baseline: Model B1, Black respondents	207.926	166	.02	.980	.04	.040	.890	16010	16285
B1.W	Baseline: Model B1, White respondents	320.091	166	.00	.976	.03	.032	.912	73958	74364
M1	Configural Invariance	528.017	332	.00	.977	.03	.033	.908	89968	90815
M2	Weak Invariance	563.259	368	.00	.977	.03	.038	.908	89931	90601
M3	Strong Invariance	591.135	383	.00	.976	.03	.040	.903	89929	90525
M4a	Strict invariance (variances)	657.701	404	.00	.970	.04	.042	.883	89954	90446
M4b	Strict invariance (variances, except *Designs II*)	633.463	403	.00	.973	.03	.041	.893	89931	90429
M4c	Strict invariance (covariances)	634.47	405	.00	.973	.03	.041	.893	89928	90416
M5	Latent variances	649.64	411	.00	.972	.03	.054	.889	89932	90389
M6	Latent mean differences of all factors	875.523	417	.00	.946	.05	.126	.798	90145	90573
M6a	Latent mean differences of Working Memory and Processing Speed constrained to be zero	653.614	413	.00	.972	.03	.055	.888	89932	90380
M7	Latent mean differences of Working Memory, Processing Speed, and *g* constrained to be zero	802.278	414	.00	.954	.04	.103	.823	90078	90521

Note. CFI: comparative fit index; RMSEA: root mean square error of approximation; SRMR: standardized root mean square residual, Mc: McDonald's non-centrality index, AIC: Akaike's information criterion, BIC: Bayesian information criterion. $n_{\text{Black}} = 180$, $n_{\text{White}} = 835$.

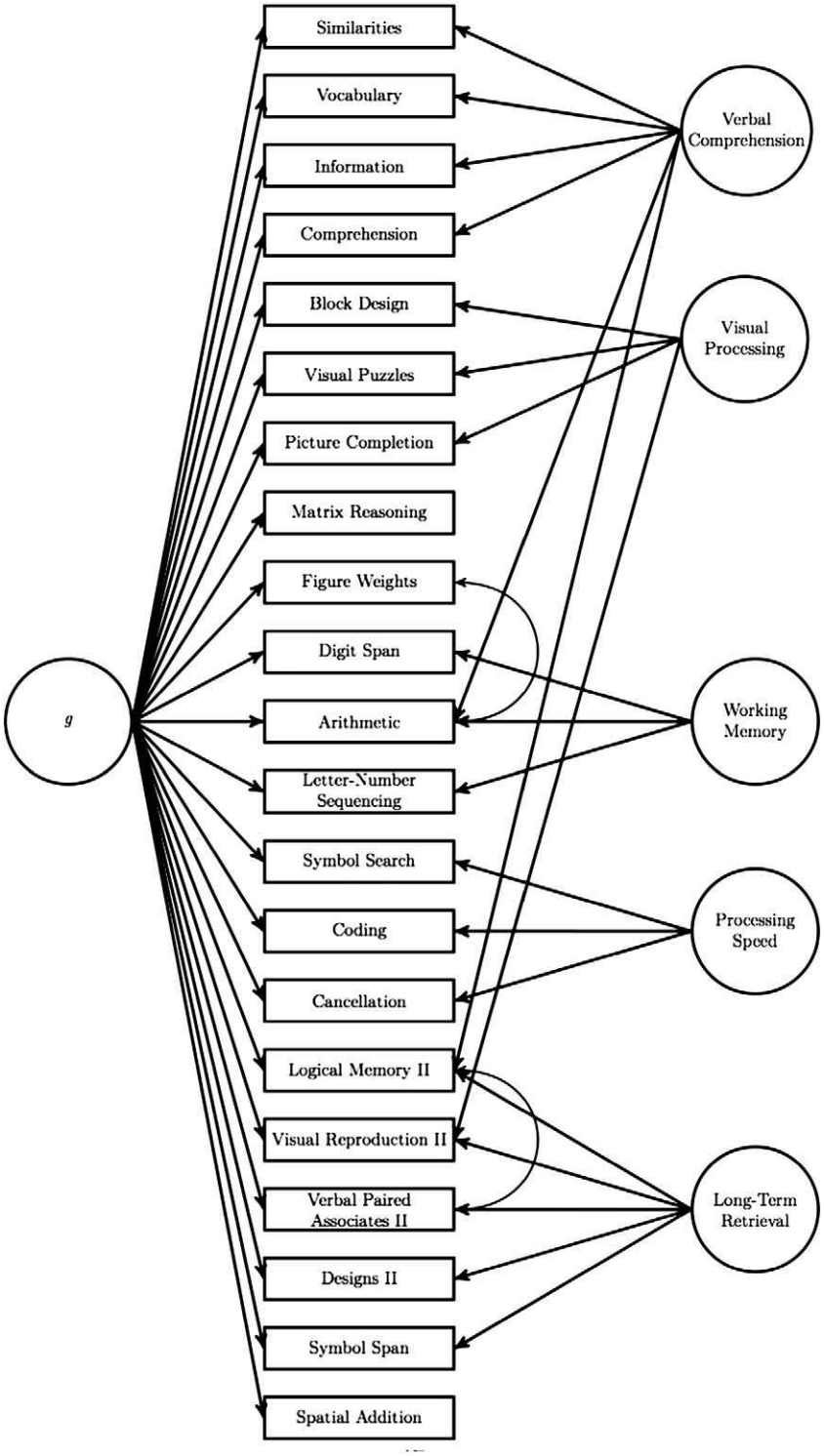


Fig. 2. Bi-factor model of the WAIS-IV and WMS-IV subtests. Subtest specific/error variance terms not shown for space considerations.

the model showed some depreciation in fit. Consequently, we examined the modification indices, which indicated that the residuals for the Designs II subtests should be freed between groups. The resulting model (M4b) fit the data better than model M4a, and only slightly worse than model M3. We then constrained the residual covariances (M4c), which did not

worsen the model fit. Thus, it appears that the construct reliabilities for Verbal Comprehension, Visual Processing, Working Memory, and Processing Speed factors are the same across groups, and almost the same for *g* and Long-Term Memory.

Next, we constrained the latent variables' variances by fixing all of them to be 1.0 (model M5). While this step is not required for assessing measurement invariance, it is required to examine if the groups used equivalent ranges of the latent variables to respond to the tests. This did not appear to worsen model fit. The factor loadings from this final model (M5) are given in Table 7.

Last, we examined the differences in the latent variable's means. First, we constrained all means to be the same (i.e., zero) across groups (M6). This model showed an appreciable depreciation in model fit. Subsequently, we examined the latent means from model M5 to see if there were any factors with minimal group differences. The latent mean differences in the Working Memory and Processing Speed factors seemed likely due to sampling error as their the 95% CIs contained zero. Consequently, we set these factors' latent means to be zero for both groups (M6a). This model fits as well as model M5. We show the latent mean differences for the factors based on model M6a in Table 8.

As the latent variables do not have an inherent mean, we set the mean of the Black group to zero and estimated the means in the White group. Thus, the values in Table 8 show how much higher (positive value) or lower (negative value) the latent mean for the White group is from the latent mean of the Black group. As the variance of the latent variables in both groups is

Table 7
Factor loadings for final model (Model M5) of WAIS-IV/WMS-IV data.

Factor	Subtest	Unstandardized estimate	SE	Standardized estimate
g				
	Similarities	1.63	0.08	0.60
	Vocabulary	1.63	0.08	0.59
	Information	1.69	0.09	0.58
	Comprehension	1.70	0.09	0.58
	Block Design	1.73	0.08	0.64
	Visual Puzzles	1.79	0.08	0.63
	Picture Completion	1.69	0.09	0.58
	Matrix Reasoning	2.03	0.08	0.71
	Figure Weights	2.11	0.09	0.73
	Digit Span	1.65	0.08	0.61
	Arithmetic	1.75	0.08	0.64
	Letter–Number Sequencing	1.70	0.09	0.58
	Symbol Search	1.53	0.08	0.54
	Coding	1.60	0.08	0.56
	Cancellation	1.13	0.10	0.40
	Logical Memory II	1.18	0.09	0.41
	Visual Reproduction II	1.52	0.10	0.50
	Verbal Paired Associates II	1.33	0.09	0.45
	Designs II	1.36	0.10	0.56
	Symbol Span	1.90	0.08	0.65
	Spatial Addition	1.94	0.10	0.65
Verbal Comprehension				
	Similarities	1.38	0.08	0.51
	Vocabulary	1.81	0.07	0.65
	Information	1.47	0.08	0.50
	Comprehension	1.61	0.08	0.55
	Arithmetic	0.49	0.08	0.18
	Logical Memory II	0.62	0.10	0.21
Visual Processing				
	Block Design	1.43	0.17	0.52
	Visual Puzzles	0.89	0.13	0.31
	Picture Completion	0.45	0.12	0.15
	Visual Reproduction II	0.48	0.14	0.16
Working Memory				
	Digit Span	2.11	0.36	0.78
	Arithmetic	0.45	0.11	0.17
	Letter–Number Sequencing	1.03	0.20	0.35
Processing Speed				
	Symbol Search	1.77	0.13	0.63
	Coding	1.51	0.12	0.53
	Cancellation	1.06	0.12	0.37
Long-Term Retrieval				
	Logical Memory II	0.64	0.14	0.22
	Visual Reproduction II	1.33	0.17	0.44
	Verbal Paired Associates II	0.96	0.14	0.32
	Designs II	1.07	0.16	0.44
	Symbol Span	0.69	0.12	0.24

Note. For all analyses, we used full information maximum likelihood estimation to account for missing data.

Table 8
Black–White mean differences on latent variables.

Factor	Estimate	SE	95% CI	
			Lower	Upper
g	1.16	0.10	0.97	1.34
Verbal Comprehension	0.23	0.11	0.01	0.45
Visual Processing	0.80	0.15	0.51	1.08
Working Memory	0	–	–	–
Processing Speed	0	–	–	–
Long-Term Retrieval	−0.35	0.16	−0.65	−0.04

Note. Estimates came from model M6a (see Table 6). Latent mean differences for Working Memory and Processing Speed were constrained to zero. For all latent variables, the variances were fixed at 1.0 and the means for the Black group were fixed at 0.0. Thus, a positive difference indicates the average score from the White group was higher, while a negative difference indicates the average score from the Black group was higher.

one, these mean differences are given in standard deviation units. The White group is approximately 1.16 SDs higher on *g*, 0.80 SDs higher on Visual Processing, and .23 SDs higher on Verbal Comprehension than the Black group. Conversely, the Black group was 0.35 SDs higher on the Long-Term Retrieval factor.

Model M6a supports the weak form of SH, so to rule out the contra hypothesis version of SH, we fit a model that allowed for mean differences only in Stratum II factors. Here, we estimated the latent means differences for Verbal Comprehension, Visual Spatial Reasoning, and Long-Term Retrieval, but constrained the latent mean differences for *g*, Working Memory, and Processing Speed to be zero (model M7). The model fit is worse than that for model M6a, indicating that subtest differences are not due to latent mean differences in Stratum II factors alone.

4. Discussion

Interpretations of the meaning of subgroup differences in average score performance on cognitive tests have been plagued by ad hoc "armchair" explanations that have sowed confusion rather than clarity among practitioners and researchers (e.g., Helms, 1997). The correlated vector (CV) method was a major step forward in establishing an empirically based method to both posit and test a coherent, parsimonious theory—called the Spearman hypothesis (SH)—that explains these differences (Jensen, 1985). The multi-group confirmatory factor analysis (MG-CFA) method represents a second step forward in providing a technique to assess measurement invariance across comparison groups, as well as provide a simultaneous test for the strong, weak and contra hypotheses associated with SH. Studies using MG-CFA have often yielded equivocal results, which we contend are primarily due to shortcomings in the way *g* has been modeled. In this article, we described how the bi-factor model (BF; Holzinger & Swineford, 1937; Jennrich & Bentler, 2011, 2012) can offer advantages to both the CV and MG-CFA approaches of examining SH.

We demonstrated the use of the BF model to examine SH in a large co-normed standardization dataset of scores from the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) and Wechsler Memory Scale-Fourth Edition (WMS-IV). This data has the advantage of including a wide variety of mental tests as well as containing an overrepresentation of memory tests, which tend to either show minimal race differences or favor Black respondents (Jensen & Reynolds, 1982; Mayfield & Reynolds, 1997). Consequently, we expected to find support for the weak form of SH in this dataset.

Our CV analysis revealed the correlation between Black–White score differences and *g* loadings to be approximately .60, with the Pearson correlation being slightly lower and the Spearman correlation being slightly higher. As there is no agreed-upon value that differentiates the strong and weak forms of SH, we are unsure if this supports the weak or strong version of SH. It is likely that these findings favor the weak version of SH because more than half the variance in the score differences are not accounted for by *g*.

The results from the MG-CFA also support the weak form of SH. While there were large mean differences in *g*, there were also substantial mean differences in the Visual Processing factor as well. In addition, there were moderate differences in the Verbal Comprehension and Long-Term Retrieval factors, with the latter favoring the Black sample. Thus, while *g* does play a part in the score differences between Black and White participants, it is not the only construct contributing to these differences.

4.1. Integration with previous literature

Our finding of large Black–White differences in *g* (1.16 SDs) and Visual Processing (0.80 SDs) is consistent with other SH studies. In Jensen's (1998) summary of SH studies, he reported the largest Black–White differences (favoring Whites) were found on tests that load highly on both *g* and a Spatial Visualization (i.e., Visual Processing) factor. More recently, Dragt's (2010) meta-analysis of SH studies confirmed Jensen's findings:

> The fact that tests that are heavily loaded on either the [Visual Processing] factor or [Short-Term Memory] factors consistently cause small deviations from the result predicted by the strong form of Spearman's hypothesis dictates that this form must be rejected. The weak form of Spearman's hypothesis, however, is strongly confirmed. (p. 61).

At the other extreme, our finding of no Black–White differences in Working Memory and a small difference favoring the Black respondents in Long-Term Retrieval is consistent with the SH literature as well (Dolan & Hamaker, 2001; Jensen & Reynolds, 1982).

As have previous studies of SH (Dolan, 2000; Dolan & Hamaker, 2001), the results from the MG-CFA indicated that there was strict invariance for the majority of the WAIS-IV/WMS-IV subtests. The only exception was the WMS-IV Designs II subtest, whose error variance was not the same between groups. Unlike previous studies, however, we were able to differentiate the effects of *g* on the group differences from the effects of the Stratum II factors. Previous MG-CFA studies that used HOF models were equivocal about whether it was differences in *g*, differences in Stratum II factors, or both that were causing the observed test score differences. Our use of a bi-factor model enabled us to show that the observed test scores were due to differences in *g* as well as differences in Stratum II factors (Visual Processing, Verbal Comprehension,

and Long-Term Retrieval). That is, our study confirmed the weak form of SH, consistent with Jensen's (1998) interpretation of the SH data.

4.2. Final thoughts on comparing approaches to investigating SH

4.2.1. Comparison of results from the current study

Our study revealed noteworthy similarities between the CV and MG-CFA approaches used to investigate the SH when using a BF model to represent *g*. Both approaches showed that *g* was estimated invariantly in both the Black and White groups as well as showed large Black–White differences on *g*. The CV analysis yielded a correlation between Black–White differences and *g* between 0.58 (Pearson) and 0.62 (Spearman), while the MG-CFA analysis yielded difference in the latent mean of *g* of 1.16 SDs.

There were some noticeable differences between the CV and MG-CFA approaches as well. First, the MG-CFA was able to uncover more nuanced information than the CV analysis. Specifically, the MG-CFA not only found differences in *g*, but also found group differences in Visual Processing (0.80 SDs favoring the White sample), Verbal Comprehension (0.23 SDs favoring the White sample), and Long-Term Retrieval factors (0.34 SDs favoring the Black sample). Second, the MG-CFA found the construct reliability estimates to be very similar between the groups for all the factors, an issue the CV method does not even attempt to address. Third, while both the CV and MG-CFA approaches showed large Black–White differences on *g*, the magnitude of the difference is somewhat larger for the CV analysis than the MG-CFA. Specifically, the *d* effect sizes that correspond to the correlations from the CV approach are 1.42 (Pearson) and 1.58 (Spearman).

4.2.2. Preferred method for assessing Spearman's hypothesis

The results from our study are in agreement with those from Dolan and his colleagues (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001) showing that the MG-CFA approach to testing SH is typically better than using CV. First, the MG-CFA approach is better able to test the assumptions inherent in SH than the CV approach. Second, by using a BF approach to using a CFA model, the approach can provide more information about the nature of the between-group differences. For example, the BF MG-CFA approach allows for an assessment of group differences in *g* and the Stratum II factors simultaneously. Thus, it allows for a direct investigation of the strong, weak, and contra forms of SH. Third, although seldom discussed in the SH literature (however, see Irwing, 2012), the BF MG-CFA approach allows for an assessment of construct reliability differences between groups, for both *g* and the Stratum II factors. The current study found strict invariance for all the subtests (except Designs II) as well as invariance in the latent variances. Thus, not only are the between-group construct reliabilities nearly identical, but both groups used equivalent ranges of the latent variables when responding to the test questions. Where strict invariance not found, however, then we could have followed the MG-CFA with an investigation of the reliability of the measured constructs (Reise, Bonifay, & Haviland, 2013).

Despite the number of benefits the MG-CFA approach has over the CV approach, the CV approach to assessing SH (or differences between any groups) is still quite common. Critiques of the CV method were issued over 15 years ago (e.g., Ashton & Lee, 2005; Dolan, 2000; Millsap, 1997), yet the method is still used. If the CV method is going to continue to be used, further work needs to be done to determine what level of the correlation between *g* and the differences in test scores is required for support of the strong vs. weak vs. contra forms of SH. The current lack of agreed-upon values has caused a variety of correlation values to be interpreted as evidence supporting *g*'s role in determining group differences (Dolan et al., 2004). A Monte Carlo study could be useful here. Specifically, after simulating data from strong, weak, and contra forms of SH, the magnitude of the correlations from a CV analysis of all the data sets could be compared to give an idea about benchmarks for support of each level of SH.

4.3. Bi-factor versus higher-order models for testing Spearman's hypothesis

All prior studies that have compared the CV and MG-CFA methods for evaluating the SH have used a higher-order factor (HOF) model. In contrast, we used a BF model and, to our knowledge, are the first to compare CV with a MG-CFA using a BF model's representation of *g* and the Stratum II factors.

If *g* were the only concern in testing SH, then it might not make much of a difference whether a BF or HOF model was used (Jensen & Weng, 1994). SH does not focus solely on *g*, however, because the weak and contra forms also considers the influence of Stratum II factors. In the HOF model, Stratum II factors are comprised of two independent components: the part that is due to *g* and the part that is independent of *g*. In the BF model, Stratum II factors are defined as constructs that influence a set of observed tests scores independent of the influence of *g* (Chen et al., 2006). Thus, Murray and Johnson (2013, p. 420) concluded, "If 'pure' measures of specific abilities are required then bi-factor model factor scores should be preferred to those from a higher-order model."

4.3.1. A bi-factor model of intelligence

Some may question whether a BF model is an appropriate representation of intelligence. HOF models have been used so often in the field and some argue that they have a stronger theoretical basis than BF models (e.g., Keith & Reynolds, 2012; Murray & Johnson, 2013). Recently, Beaujean (submitted for publication) argued that a BF theory of intelligence does exist—the one that started with Spearman's conceptualizations of *g*, group factors, and specific factors, and then evolved in Carroll's three-stratum theory.

First, a BF model's representation of *g* is consistent with Spearman's conceptualization because the BF model is just an extension of Spearman's two-factor theory that allows for Stratum II (group) factors (Holzinger & Swineford, 1939). This is not surprising, given Holzinger's close association with Spearman (Harman, 1954). Moreover, Spearman's conceptualization of group factors is aligned with the BF model (Spearman, 1933) and he accepted the *g* factor extracted from a BF model to be the same as that from his two-factor theory (Spearman, 1946).

Second, John Carroll's conceptualization of intelligence is more consistent with a BF model than a HOF model. Carroll (1997) argued that *g* should be extracted from a set of cognitive ability measures first, and then the Stratum II factors should be

formed from covariances residualized after extracting *g*. This is the same idea Holzinger and Swineford (1937) used in developing the BF model.

While Carroll (1996) often presented his three-stratum theory as a higher-order model in figures, he warns against taking the structure of his figures "too literally or precisely" (p. 4) because he explicitly preferred the BF model to the HOF model. This is most noticeable in the CFAs Carroll conducted in order to verify his three-stratum model, as he consistently chose to use BF models instead of HOF models (Carroll, 1997, 1995).

One may argue that a HOF model is more preferable to a BF model because *g* is best thought of as an abstraction of Stratum II factors, not a direct influence on tests. This argument not only contradicts Carroll's (1996) conceptualization of *g*, but also is contrary to Spearman's initial conceptualization of *g* as having direct influences on the measured tests (Hart & Spearman, 1912).

5. Conclusion

The CV method was a major contribution to the study of SH. The HOF MG-CFA method improved the CV method by providing a technique to examine the assumptions underlying the use of CV. We believe that the BF MG-CFA approach makes an additional contribution to the field of studying SH because it can provide a clearer picture of the contributions of *g* and Stratum II factors to the differential size of group differences.

References

Andersen, R. (2012). Methods for detecting badly behaved data: Distributions, linear models, and beyond. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology. Data analysis and research publication*, *3*. (pp. 5–26). Washington, DC: American Psychological Association.

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, *33*, 431–444. http://dx.doi.org/10.1016/j.intell.2004.12.004.

Beaujean, A. A. (2014a). *Latent variable modeling using R: A step-by-step guide*. New York, NY: Routledge.

Beaujean, A. A. (2014b). ***R*** *syntax to accompany Best Practices in Exploratory Factor Analysis (2014) by Jason Osborne*. Waco, TX: Baylor Psychometric Laboratory (Retrieved from http://www.jwosborne.com).

Beaujean, A. A. (2015n). *A bi-factor theory of intelligence does exist.* (Manuscript submitted for publication).

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell–Horn–Carroll models for predicting language achievement: Differences between bi-factor and higher-order factor models. *Psychological Assessment*, *26*, 789–805. http://dx.doi.org/10.1037/a0036745.

Benson, N., Beaujean, A. A., & Taub, G. E. (2015). Using score equating and measurement invariance to examine the Flynn effect in the Wechsler adult intelligence scale. *Multivariate Behavioral Research* (in press).

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. http://dx.doi.org/10.1037/a0017767.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*, 461–483. http://dx.doi.org/10.1207/s15328007sem0703_6.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235) (2nd ed.). New York, NY: Russell Sage Foundation.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. http://dx.doi.org/10.1207/s15327906mbr3601_05.

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

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. http://dx.doi.org/10.1207/s15327906mbr3003_6.

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

Carroll, J. B. (1997). Theoretical and technical issues in identifying a factor of general intelligence. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to the bell curve* (pp. 125–156). New York, NY: Springer-Verlag.

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

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

Colom, R., & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12–18 year olds. *Personality and Individual Differences*, *36*, 75–82. http://dx.doi.org/10.1016/S0191-8869(03)00053-9.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research*, *35*, 21–50. http://dx.doi.org/10.1207/s15327906mbr3501_2.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating black-white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC, and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychological research*, *Vol. 6*. (pp. 31–60). Huntington, NY: Nova Science Publishers.

Dolan, C. V., & Lubke, G. H. (2001). Viewing Spearman's hypothesis from the perspective of Multigroup PCA: A comment on Schonemann's criticism. *Intelligence*, *29*, 231–245. http://dx.doi.org/10.1016/s0160-2896(00)00054-4.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence*, *32*, 155–173. http://dx.doi.org/10.1016/j.intell.2003.09.001.

Dragt, J. (2010). *Causes of group differences studied with the method of correlated vectors: A psychometric meta-analysis of Spearman's hypothesis.* (Master's thesis) Amsterdam, the Netherlands: University of Amsterdam.

Eells, K. (1951). *Intelligence and cultural differences: A study of cultural learning and problem solving*. Chicago, IL: University of Chicago.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430–457. http://dx.doi.org/10.1207/S15328007SEM0803_5.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*, 477–531.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, *42*, 37–48. http://dx.doi.org/10.1016/j.paid.2006.06.019.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth 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. http://dx.doi.org/10.1080/00273171.2013.804398.

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

Gottfredson, L. S. (2005). Implications of cognitive differences for schooling within diverse societies. In C. C. L. Frisby, & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology*. New York: Wiley.

Gustafsson, J. -E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research*, *27*, 239–247. http://dx.doi.org/10.1207/s15327906mbr2702_7.

Harman, H. H. (1954). Karl John Holzinger. *Psychometrika*, *19*, 95–96. http://dx.doi.org/10.1007/BF02289158.

Harrington, D. (2009). *Confirmatory factor analysis*. New York, NY: Oxford University Press.

Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology*, *5*, 51–84. http://dx.doi.org/10.1111/j.2044-8295.1912.tb00055.x.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. http://dx.doi.org/10.3102/10769986006002107.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist*, *47*, 1083–1101. http://dx.doi.org/10.1037/0003-066X.47.9.1083.

Helms, J. E. (1997). The triple quandary of race, culture, and social class in standardized cognitive ability testing. In D. P. Flanagan, J. L. Genshaft, & P. J. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 517–532). New York, NY: Guilford Press.

Helms-Lorenz, M., Van de Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence*, *31*, 9–29. http://dx.doi.org/10.1016/S0160-2896(02)00111-3.

Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., & Salthouse, T. A. (2011). Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment*, *18*, 178–191. http://dx.doi.org/10.1177/1073191110393106.

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

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution. Supplementary Educational Monographs*, *48*, Chicago, IL: University of Chicago Press.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. http://dx.doi.org/10.1007/bf02289447.

Horn, J. (1997). On the mathematical relationship between factor or component coefficients and differences between means. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *16*, 721–728.

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

Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences*, *53*, 126–131. http://dx.doi.org/10.1016/j.paid.2011.05.001.

Jennrich, R. I., & Bentler, P. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537–549. http://dx.doi.org/10.1007/s11336-011-9218-4.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*, 442–454. http://dx.doi.org/10.1007/s11336-012-9269-1.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Jensen, A. R. (1984). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures*, *7*, 93–118. http://dx.doi.org/10.1016/s0140-1750(84)80001-9.

Jensen, A. R. (1985). The nature of the Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*, 193–263. http://dx.doi.org/10.1017/s0140525x00020392.

Jensen, A. R. (1992). Spearman's hypothesis: Methodology and evidence. *Multivariate Behavioral Research*, *27*, 225–233. http://dx.doi.org/10.1207/s15327906mbr2702_5.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger Publishers/Greenwood.

Jensen, A. R. (2001). Spearman's hypothesis. In J. M. Collis, & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 3–24). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Jensen, A. R., & Figueroa, R. A. (1975). Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology*, *67*, 882–893. http://dx.doi.org/10.1037/0022-0663.67.6.882.

Jensen, A. R., & Osborne, R. T. (1979). Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison. *Indian Journal of Psychology*, *54*, 75–87.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423–438. http://dx.doi.org/10.1016/0191-8869(82)90007-1.

Jensen, A. R., & Weng, L. -J. (1994). What is a good g? *Intelligence*, *18*, 231–258. http://dx.doi.org/10.1016/0160-2896(94)90029-9.

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

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. A. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York, NY: Guilford.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies* (pp. 121–147). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*, 57–64. http://dx.doi.org/10.1027/1614-2241.2.2.57.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research*, *36*, 299–324. http://dx.doi.org/10.1207/S15327906299-324.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology*, *56*, 231–248. http://dx.doi.org/10.1348/000711003770480020.

Lupi, M. H., & Ting Woo, J. Y. (1989). Issues in the assessment of East Asian handicapped students. *Assessment for Effective Intervention*, *14*, 147–158. http://dx.doi.org/10.1177/153450848901400301.

Lynn, R., & Owen, K. (1994). Spearman's hypothesis and test score differences between whites, Indians, and blacks in South Africa. *Journal of General Psychology*, *121*, 27–36.

Mardia, K. (1980). Tests of univariate and multivariate normality. In P. R. Krishnaiah (Ed.), *Handbook of statistics*, *Vol.1*. (pp. 279–320). New York, NY: North Holland.

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

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

Mayfield, J. W., & Reynolds, C. R. (1997). Black–White differences in memory test performance among children and adolescents. *Archives of Clinical Neuropsychology*, *12*, 111–122. http://dx.doi.org/10.1016/S0887-6177(96)00016-9.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, *29*, 409–454. http://dx.doi.org/10.1207/s15327906mbr2904_5.

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

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568–592. http://dx.doi.org/10.1037/0021-9010.93.3.568.

Miller, D. I., Davidson, P. S. R., Schindler, D., & Messier, C. (2013). Confirmatory factor analysis of the WAIS-IV and WMS-IV in older adults. *Journal of Psychoeducational Assessment*, *31*, 375–390. http://dx.doi.org/10.1177/0734282912467961.

Millis, S. R., Malina, A. C., Bowers, D. A., & Ricker, J. H. (1999). Confirmatory factor analysis of the Wechsler memory scale-III. *Journal of Clinical and Experimental Neuropsychology*, *21*, 87–93. http://dx.doi.org/10.1076/jcen.21.1.87.937.

Millsap, R. E. (1997). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *16*, 750–757.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mulaik, S. A. (1992). Guttman's "last paper" [Special issue]. *Multivariate Behavioral Research*, *27*(1).

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal*, *4*, 193–211. http://dx.doi.org/10.1080/10705519709540071.

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. http://dx.doi.org/10.1016/j.intell.2013.06.004.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of Black–White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21–43. http://dx.doi.org/10.1016/0160-2896(87)90024-9.

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. http://dx.doi.org/10.1037/a0032086.

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

Nyborg, H., & Jensen, A. R. (2000). Black–White differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences*, *28*, 593–599. http://dx.doi.org/10.1016/s0191-8869(99)00122-1.

Price, L. R., Tulsky, D., Millis, S., & Weiss, L. (2002). Redefining the factor structure of the Wechsler memory scale-III: Confirmatory factor analysis with cross-validation. *Journal of Clinical and Experimental Neuropsychology*, *24*, 574–585. http://dx.doi.org/10.1076/jcen.24.5.574.1013.

R. Development Core Team (2014). *R: A language and environment for statistical computing. (Version 3.01) [Computer Program]*. Vienna, Austria: R Foundation for Statistical Computing.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*, 299–331. http://dx.doi.org/10.1016/s0005-7894(04)80041-8.

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

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

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559. http://dx.doi.org/10.1080/00223891.2010.496477.

Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, *71*, 684–711. http://dx.doi.org/10.1177/0013164410378690.

Revelle, W. (2012). *psych: Procedures for psychological, psychometric, and personality research. (Version 1.4.8) [Computer Program].* Evanston, IL: Northwestern University.

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

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

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297–330. http://dx.doi.org/10.1111/j.1744-6570.2001.tb00094.x.

Rushton, J. P. (1998). The "Jensen effect" and the "Spearman–Jensen hypothesis" of Black White IQ differences. *Intelligence*, *26*, 217–225. http://dx.doi.org/10.1016/s0160-2896(99)80004-x.

Rushton, J. P. (2003). Race differences in g and the "Jensen effect.". In H. Nyborg (Ed.), *The scientific study of general intelligence: A tribute to Arthur R. Jensen* (pp. 147–186). Amsterdam: Pergamon.

Rushton, J. P., & Jensen, A. R. (2003). African–White IQ differences from Zimbabwe on the Wechsler intelligence scale for children-revised are mainly on the g factor. *Personality and Individual Differences*, *34*, 177–183. http://dx.doi.org/10.1016/s0191-8869(02)00024-7.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235–294. http://dx.doi.org/10.1037/1076-8971.11.2.235.

Salthouse, T. A. (2009). Decomposing age correlations on neuropsychological and cognitive variables. *Journal of the International Neuropsychological Society*, *15*, 650–661. http://dx.doi.org/10.1017/S1355617709990385.

Schmid, J., Jr. (1957). The comparability of the bi-factor and second-order factor patterns. *The Journal of Experimental Education*, *25*, 249–253. http://dx.doi.org/10.2307/20154047.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61. http://dx.doi.org/10.1007/bf02289209.

Schmiedek, F., & Li, S. -C. (2004). Toward an alternative representation for disentangling age-associated differences in general and specific cognitive abilities. *Psychology and Aging*, *19*, 40–56. http://dx.doi.org/10.1037/0882-7974.19.1.40.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 99–144) (3rd ed.). New York, NY: Guilford.

Schönemann, P. H. (1997). Famous artefacts: Spearman's hypothesis. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *16*, 665–694.

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

Spearman, C. E. (1927). *The abilities of man: Their nature and measurement.* New York, NY: Blackburn Press.

Spearman, C. (1933). The factor theory and its troubles. III. Misrepresentation of the theory. *Journal of Educational Psychology*, *24*, 591–601. http://dx.doi.org/10.1037/h0073929.

Spearman, C. E. (1946). Theory of general factor. *British Journal of Psychology*, *36*, 117–131. http://dx.doi.org/10.1111/j.2044-8295.1946.tb01114.x.

te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology*, *82*, 675–687. http://dx.doi.org/10.1037/0021-9010.82.5.675.

te Nijenhuis, J., & van der Flier, H. (2003). Immigrant–majority group differences in cognitive performance: Jensen effects, cultural effects, or both? *Intelligence*, *31*, 443–459. http://dx.doi.org/10.1016/S0160-2896(03)00027-8.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327. http://dx.doi.org/10.1007/bf02293557.

Ward, L. C., Bergman, M. A., & Hebert, K. R. (2012). WAIS-IV subtest covariance structure: Conceptual and statistical considerations. *Psychological Assessment*, *24*, 328–340. http://dx.doi.org/10.1037/a0025614.

Wechsler, D. (2008a). *Wechsler adult intelligence scale* (4th ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008b). *Wechsler adult intelligence scale: Technical and interpretive manual* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2009). *Wechsler memory scale* (4th ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D., Holdnack, J. A., & Drozdick, L. W. (2009). *Wechsler memory scale-fourth edition technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013a). WAIS-IV and clinical validation of the four- and five-factor Interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 94–113. http://dx.doi.org/10.1177/0734282913478030.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013b). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 114–131. http://dx.doi.org/10.1177/0734282913478032.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. A. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Newbury Park, CA: Sage.

White, J. (1984). *The psychology of blacks.* Englewood Cliffs, NJ: Prentice-Hall.

Woodley, M. A., te Nijenhuis, J., Must, O., & Must, A. (2014). Controlling for increased guessing enhances the independence of the Flynn effect from g: The return of the Brand effect. *Intelligence*, *43*, 27–34.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. (Unpublished doctoral dissertation). University of California-Los Angeles, Los Angeles, California.

Yung, Y. -F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128. http://dx.doi.org/10.1007/bf02294531.