Intelligence 52 (2015) 71–79

Contents lists available at ScienceDirect

Intelligence

Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of *g*

Gilles E. Gignac

School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 23 May 2015
Received in revised form 19 July 2015
Available online xxxx

Keywords:
Raven's Progressive Matrices
General intelligence
Fluid intelligence

ABSTRACT

It has been claimed that Raven's Progressive Matrices is a pure indicator of general intelligence (*g*). Such a claim implies three observations: (1) Raven's has a remarkably high association with *g*; (2) Raven's does not share variance with a group-level factor; and (3) Raven's is associated with virtually no test specificity. The existing factor analytic research relevant to Raven's and *g* is very mixed, likely because of the variety of factor analytic techniques employed, as well as the small sample sizes upon which the analyses have been performed. Consequently, the purpose of this investigation was to estimate the association between Raven's and g, Raven's and a theoretically congruent group-level factor, and Raven's test specificity within the context of a bifactor model. Across several large samples, it was observed that Raven's (1) shared approximately 50% of its variance with *g*; (2) shared approximately 10% of its variance with a fluid intelligence group-level factor orthogonal to *g*; and (3) was associated with approximately 25% test specific reliable variance. Overall, the results are interpreted to suggest that Raven's is not a particularly remarkable test with respect to *g*. Potential implications relevant to the commonly articulated central role of Raven's in *g* factor theory, as well as the Flynn effect, are discussed. Finally, researchers are discouraged to include only Raven's in an investigation, if a valid estimate of *g* is sought. Instead, as just one example, a four-subtest combination from the Wechsler scales with a *g* validity coefficient of .93 and 14 min administration time is suggested.

1. Introduction

Raven's Progressive Matrices (RPM; Raven, 1940; Raven & Court, 1998) have been contended in the intelligence literature to be a pure (or nearly so) indicator of Spearman's (1927) general intelligence (*g*; Deary & Smith, 2004; Eysenck, 1998; Jensen, 1998; Llabre, 1984; Neisser, 1998; Thorndike, 1986; Vernon, 1947). Correspondingly, researchers across various disciplines within psychology commonly include a single measure of intelligence in their design, Raven's, and interpret their results (or lack thereof) under the pretence that they have measured *g* (e.g., Basso, De Renzi, Faglioni, Scotti, & Spinnler, 1973; Corben et al., 2006; Day et al., 2005; Schellenberg & Moreno, 2010; Walker, Pierre, Christie, & Chang, 2013). The purpose of this investigation was to review critically the literature upon which claims of "pure *g*" have been made. Additionally, the question will be examined empirically across several data sets, in conjunction with a rigorous, direct, factor analytic approach, the bifactor model (Gustafsson & Balke, 1993).

2. Raven's and *g*: background

Although there are several versions of RPM that have been published over the years (Raven, 1940; Raven, 1966; Raven, Court, & Raven, 1994; Raven, Raven, & Court, 1962), they all have in common the inclusion of items that consist of visually presented figural matrices within which one piece is missing. In order to complete a particular matrix, the participant must choose one piece amongst several alternatives which are presented below the matrix. In more practical terms, this would involve the successful identification of one or more figural patterns associated with the displayed pieces within the matrix.

It is commonly stated that Raven's may be considered the purest expression of *g* (Martinez, 2013). Spearman (1946) considered matrices type tests to be the nearest representation of *g*, as they contained items that required the eduction of relations and correlates to solve, a view shared by others (e.g., Holyoak, 2012; Jensen, 1998; Thorndike, 1986). Some researchers view the *g* factor saturation associated with Raven's to be so substantial that the item-total correlations may be considered an accurate estimate of each items *g* loading (Rushton & Skuy, 2001). However, few of these sources provide primary references, or, in many cases, any references at all, to support the contention that Raven's is a pure measure of *g*.

Perhaps one of the researchers who has most frequently asserted Raven's to be a pure measure of *g* is Arthur Jensen (e.g., Jensen, 1973; Jensen, 1980a; Jensen, 1987; Jensen, 1998). However, again, only occasionally has a reference been provided to support such a claim. In a review of Raven's, Jensen (1980a, p. 646) wrote:

> "Factorially the Progressive Matrices apparently measures *g* and little else (Burke, 1958). The loadings that are occasionally found on other

E-mail address: gilles.gignac@uwa.edu.au.

http://dx.doi.org/10.1016/j.intell.2015.07.006
0160-2896/© 2015 Elsevier Inc. All rights reserved.

> "perceptual" and "performance"-type factors, independently of *g*, are usually so trivial and inconsistent from one analysis to another as to suggest that the RPM does not reliably measure anything but *g* in the general population."

Thus, Jensen (1980a) relied upon Burke (1958) for evidence to support his assertion. Although it is the case that Burke (1958) did review some factor analytic research to suggest that Raven's may be a relatively pure measure of *g*, it is also the case that Burke (1958) reviewed a substantial amount of factor analytic research to suggest that Raven's was not a relatively pure measure of *g*. Importantly, Burke (1958, p. 212) concluded the following on the matter:

> "The evidence is not convincing that [Progressive Matrices] has validity as a pure measure of the Spearman construct of *g*; and doubt may be raised whether such a construct can be measured independently of the modality through which it is expressed, the selectivity of the subjects and their sex, and possibly the presuppositions of the factor analyst."

In another investigation, Jensen (1987) cited Vernon (1983) as support for the contention that Raven's may be considered a relatively pure indicator of *g*. Vernon (1983) tested 100 university students on reaction time measures, as well as the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) and the Advanced Progressive Matrices (APM; Raven, 1966). Based on the extraction of a single principal component, Vernon (1983) reported the APM to be associated with a *g* loading of .797, which was numerically larger than the next largest *g* component loading (Block Design $g = .692$). Thus, based on the reported APM component loading, it would appear that Raven's is a relatively substantial indicator of *g*. However, as Vernon (1983) reported the results for a single component only, it was not possible to evaluate the possibility of a secondary Raven's loading. Finally, the sample size of 100 would not be considered large for the purposes of estimating a component loading in a precise manner.

In contrast to the numerous authors cited above, there are some who have expressed a more uncertain view in relation to the possibility of Raven's as a pure indicator of *g*. For example, Carroll (1993) acknowledged Raven's as a good measure of *g*, but it was not identified as clearly the single best measure of *g*. Furthermore, Carroll (1993) suggested that there was some evidence that Raven's loaded on a separate induction intelligence factor or possibly a separate spatial ability factor. Brody (1987) was also unconvinced by the position that Raven's was a pure measure of *g*. Instead, Brody (1987) suggested that previously reported factor loadings of intelligence test batteries appear to be too highly variable to allow any meaningful conclusions in that respect. Such a position suggests that no test has an intrinsic *g* loading, as it would be expected to fluctuate based on the other tests included in the analysis. Also, the fact that the Flynn effect is most pronounced on Raven's may be suggested to be evidence contrary to the notion that it is a pure measure of *g* (Johnson, 2012). Finally, in a discussion relevant to test development, Sidney Irvine opined that Raven's could not possibly be a pure measure of *g*, as it would be expected to be associated with a large amount of test specific variance (Wainer, 2002).

3. Some empirical research since Burke (1958)

Brody's (1987) point in relation to the substantial amount of variability in the factor analytic results is underscored by a review of some of the factor analytic literature published since Burke (1958). That is, some researchers report Raven's to be the most substantial indicator of *g*, while others do not. Similarly, Raven's is occasionally observed to load onto a secondary group-factor, and on other occasions it is not.

For example, Rogers, Fisk, and Hertzog (1994) administered a battery of 20 cognitive ability tests, including the Standard Progressive Matrices (SPM; Raven, Court, & Raven, 1977), to a sample of 140 participants (mix of university students and members of the community) in an investigation relevant to intelligence and visual search performance across age. The battery consisted of four semantic knowledge tests (Vocabulary, Analogies, Information, Controlled Associations), three induction tests (Mathematic Reasoning, SPM, Letter Sets), three working memory tests (Computation Span, Listening Span, Alphabet Span), four perceptual speed tests (Digit Symbol, Identical Pictures, Number Comparison, Finding *A*s), three semantic memory tests (Semantic Matching, Lexical Access, Synonym Matching), and three psychomotor speed tests (Simple Reaction Time, Making *X*s, Crossing Lines). Based on a higher-order model, the SPM was observed to have an association of .76 with *g* (Schmid–Leiman transformed) and no other group-level factor, which would suggest that Raven's may be exclusively associated with *g*.

In another investigation, DeYoung, Peterson, and Higgins (2005) administered Vocabulary, Similarities, Arithmetic, Digit Symbol, and Block Design from the WAIS-III, as well as the APM (Raven, Raven & Court, 1998), to a sample of 174 university students. Based on the extraction of a single factor, the APM was reported to be associated with the second largest *g* loading (.72). The largest *g* loading was associated with Block Design (.74). The inter-subtest correlation between the APM and Block Design was the largest at $r = .65$, however, there were no other appropriate subtests with which to model a possible a group-level spatial reasoning factor.

Kranzler and Jensen (1991) administered the WAIS and the APM (Raven, 1966) to a sample of 101 university students. Based on a Schmid–Leiman transformation of the higher-order model solution, the APM was reported to have a rather small *g* loading of .44, which was numerically lower than the mean subtest *g* loading of .48. Furthermore, the APM was also observed to load (.48) onto a Perceptual Organisation group-level factor. These results would suggest that Raven's is neither a remarkable *g* loading test, nor exclusively associated with *g*.

Finally, Ackerman (1988) administered a battery of 22 cognitive ability tests (including the SPM) to a sample of 65 undergraduate students and performed a higher-order model factor analysis with a corresponding Schmid–Leiman transformation. The battery consisted of five tests of perceptual speed (Name Comparison, Clerical Ability, Perceptual Speed, Letter/Number Substitution, Scattered Xs), three tests of movement speed (Circle Tapping, Square Marking, Pursuit Aiming), three tests of memory (Object–Number, Picture Number, First and Last Names), three tests of verbal ability (Analogies, Vocabulary, Word Beginnings), five tests of reasoning (SPM, Patterns, Number Series, Figure Classification, Letter Sets), and three indicators of reaction time (simple, two-choice, and four-choice). The SPM was observed to be associated with a *g* loading of .57, which was numerically smaller than several other tests (Letter Sets $g = .62$; Letter/Number Substitution $g = .60$). Thus, Raven's was not observed to be associated with the largest *g* loading in this investigation. Furthermore, the SPM was reported to be associated with a secondary loading of .45 on the group-level reasoning factor, which was similar in magnitude to the secondary loadings associated with the other four reasoning tests. Thus, Raven's was neither observed to be associated with a remarkably large *g* loading, nor was it observed to be exclusively associated with *g*.

Although the above is not an exhaustive review of the relevant factor analytic literature published since Burke (1958), it may nonetheless be plausible to suggest that the published results in relation to Raven's *g* factor saturation are highly variable. Arguably, the very mixed results may be due to the rather small sample sizes (often $N \leq 100$) upon which the factor analyses have been performed. Additionally, the different factor analytic techniques that were employed may be expected to yield variation in the results, not to mention results of questionable validity, in some cases. For example, a non-negligible percentage of the investigations simply extracted a single component to estimate the loading of Raven's on *g*, which is arguably an inadequate approach to the estimation a non-biased general factor solution (Ashton, Lee, & Vernon, 2001; Gignac, 2006a). Also, the extraction of a single component necessarily precludes the possibility of a secondary Raven's group-level factor loading. Additionally, the number and nature of subtests included in the analyses may have been arguably insufficient to yield a relevant visual–spatial intelligence (g_v) or fluid intelligence (g_f) group-level factor. Finally, a large percentage of the investigations were based on university students, which would limit the generalisability of the results to the population.

Arguably, a more rigorous approach to the evaluation of Raven's *g* factor saturation would take advantage of the benefits of structural equation modeling. For example, in a paper not specifically relevant to Raven's *g* saturation, Gignac (2008) modeled a bifactor model ($N = 198$) with one first-order *g* factor (defined by 12 subtests, one of which was the APM) and four nested group-level factors, g_f, crystallised intelligence (g_c), short-term memory (g_{sm}), and processing speed (g_s). The bifactor model may be considered especially useful in this case, as it models the association between all subtests and the *g* factor directly, as well as the simultaneous association between all subtests and a particular group-factor directly (Gustafsson & Balke, 1993). Gignac (2008) reported the APM to be associated with a *g* loading of .55 and a nested g_f group-level factor loading of .36. Notably, the mean *g* loading across all 12 subtests was .46. Consequently, the APM was not observed to be a remarkably strong indicator of *g* (in fact, two subtests loaded more strongly). Furthermore, the two subtests that defined the nested g_f factor alongside the APM were associated with loadings of .29 and .38; thus, very comparable to the APM's secondary g_f factor loading. Based on the results of Gignac (2008), it would be difficult to consider Raven's a noteworthy test with respect to *g*.

Although the sample size upon which Gignac's (2008) analyses were performed was somewhat respectable ($N = 198$), the participants were university students (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004), which would be expected to be associated with some level of range restriction in intelligence test scores. Consequently, the results may not be generalizable to the adult population. Unfortunately, Gignac (2008) appears to be the only factor analytic investigation to have modeled the simultaneous unique direct associations between Raven's, *g*, and a group-level factor.

4. Group-level factors: their nature and measurement

It may be suggested that there is a moderate amount of consensus on the presence of approximately 10 cognitive ability dimensions (e.g., g_f, g_c, g_v, and g_{sm}), in addition to *g*, within the domain of conventional intelligence testing (Carroll, 2003). These relatively narrow cognitive ability dimensions are often referred to as 'group-level factors' within the factor analytic research, so as to distinguish them from the general factor. Whether estimated from a higher-order modeling or bifactor modeling framework, the strength of the group-level factors tend to be relatively weak, in comparison to the *g* factor in intelligence research (e.g., Gignac & Watkins, 2013). However, they remain an active area of research with respect to their nature and incremental predictive validity beyond *g*.

For example, based on a bifactor model of the WAIS and an inspection time indicator, Crawford, Deary, Allan, and Gustafsson (1998) found that inspection time had unique effects on both a nested perceptual organisation factor (−.39), as well as *g* (−.19). Similarly, Brunner (2008) found that socioeconomic status was related uniquely to both a nested verbal ability factor (.11) and *g* (.41). Thus, as the group-level factors predict theoretically congruent criteria, they are likely associated with a certain amount of substantive variance. However, it should also be acknowledged that a portion of the group-level factor variance may also be shared method variance (Lubinski & Dawis, 1992). With respect to this investigation, as Raven's is a measure of fluid intelligence (Wilhoit & McCallum, 2003), and the item stimuli are visual/spatial in nature, it may be contended that Raven's should evidence a secondary group-level factor loading on a g_f/g_v type group-level factor, whether do to shared substantive variance, shared method variance, or both.

The two most common methods used to model group-level factors and a general factor are the higher-order model and the bifactor model (Rindskopf & Rose, 1988). Although debate surrounds preferences for each particular model (Canivez, in press; Gignac, 2008), as well as why one model may fit better than the other (Morgan, Hodge, Wells, & Watkins, 2015; Murray & Johnson, 2013), one indisputable difference between the two models is that the higher-order model imposes a within group-level factor proportionality constraint on the strength of the association between each subtest and the corresponding general factor and the respective group-level factor residual (Schmiedek & Li, 2004). Consequently, the factor loadings derived from a higher-order model tend to be more homogenous, arguably unnaturally so, than the corresponding bifactor model loadings (Brunner, 2008; Gignac & Watkins, 2013). From a more practical perspective, an advantage associated with the bifactor model is that the associations between the subtests and the factors are estimated directly, as are the corresponding standard errors and corresponding *p* values (Gignac, 2007). By contrast, in the higher-order model, the factor solution must be first submitted to a Schmid–Leiman transformation in order to estimate the associations between the subtests and the factors, as well as the corresponding standard errors. Despite the above, although seemingly very dissimilar models, the higher-order model and the bifactor model do share a number of important features and can be expected to yield similar results in many cases (Gignac, 2008).

In light of the above, the purpose of this investigation was to evaluate the *g* factor saturation associated with Raven's based on relatively representative and relatively large sample sizes ($N \geq 200$), as well as a rigorous, direct data analytic approach: the bifactor model. More specifically, the main purpose of this investigation was to evaluate the contention that Raven's is a pure (or nearly so) indicator of *g*. Such a contention would imply that Raven's is (1) associated with a remarkably large loading on *g*; (2) not associated with a secondary group-level factor loading, and (3) not associated with a substantial amount of test specificity.

5. Method

5.1. Samples

The data analysed in this investigation were derived from three well-known samples: (1) the Marshalek, Lohman, and Snow (1983) sample of 241 Palo Alto, California high-school students[1]; (2) the Minnesota Study of Twins Reared Apart sample (MISTRA; Johnson & Bouchard, 2011) of 433 adults (56.9% female) with a mean age of 42.7 years (combination of twin pairs, family members, and friends; age range: 18–79; see Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004, for further details); and (3) the Gustafsson (1984) sample of 981 based on 6th grade (12 years old) school children (49% boys). In all three cases, the sample correlation matrices were used as input for the analyses. To help avoid confusion, it will be noted here that two separate correlation matrices (and models) were ultimately derived from the MISTRA sample. The first MISTRA correlation matrix was based on the Comprehensive Ability Battery (Hakstian & Cattell, 1975) subtests in addition to the SPM (Raven, 1941). The second MISTRA correlation matrix was based on the WAIS subtests in addition to (the same) SPM. Thus, in total, four sample correlation matrices were used in this investigation to evaluate the hypotheses.

5.2. Measures

The Marshalek et al. (1983) study included a total of 34 tests. However, for the purposes of this investigation, only 19 of the subtests were selected for analysis for three reasons. First, several of the subtests were observed to be rather poor indicators of general intelligence (e.g., Street Gestalt and Harshman Gestalt). Secondly, several subtests were difficult

[1] Very few details were supplied by Marshalek et al. (1983) with respect to the sample. Instead, they cite Snow, Lohman, Marshalek, Yalow, and Webb (1977) for further information. Unfortunately, this document is an unpublished technical report. The correlation matrix used in this investigation was obtained from a report prepared by Meng (2005).

to categorise with respect to a group-level factor (e.g., Camouflage Words). Finally, 11 of the subtests were derived from the WAIS, which was considered somewhat redundant, as the WAIS was also administered in the MISTRA study. Given the above, and the descriptions and results reported in Marshalek et al., four subtests were selected as indicators of g_v (Advanced Progressive Matrices, Surface Development, Paper Form Board, Paper Folding), four as indicators of g_s (Identical Pictures, Find As, Number Comparison, Digit Symbol), four as indicators of g_q (Necessary Arithmetic Operations, Arithmetic Computation, Arithmetic Concepts, Arithmetic Applications), four indicators of g_c (Terman Concept Mastery, Reading Comprehension, Reading Vocabulary, Vocabulary) and three indicators of g_{sm} (Auditory Letter Span, Visual Number Span, Digit Span).

Participants in the MISTRA study were administered a total of three intelligence batteries in addition to the SPM (42 subtests in total). For the purposes of this investigation, Raven's was examined separately with respect to the factors associated with the Comprehensive Ability Battery (CAB; Hakstian & Cattell, 1975) and with respect to the factors associated with the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955).[2] The CAB consists of 14 subtests, three of which may be classified as indicators of g_c (Spelling, Proverbs, and Vocabulary), three indicators of g_{sm} (Meaningful Memory, Associative Memory, Memory Span), three indicators of g_s (Closure, Fluency, Perceptual Speed), five indicators of g_f (Standard Progressive Matrices, Inductive Reasoning, Flexibility of Closure, Mechanical Ability, Spatial Ability), and one indicator of quantitative reasoning (Numerical Ability).

The second battery in the MISTRA data that was examined, the WAIS (Wechsler, 1955), consists of 11 subtests, four of which are known (see Cohen, 1957) to measure a g_c factor (Vocabulary, Information, Comprehension, Similarities), four a g_v factor (Raven, Block Design, Picture Completion, Object Assembly, Picture Arrangement), and three a Freedom from Distractibility factor (Digit Span, Arithmetic, Digit Symbol).[3]

Gustafsson (1984) administered a total of 16 tests, however, four of those tests were split into odd and even items to help increase the number of indicators per group-level factor included in his CFA model. Those four tests were Mental Folding, Card Rotation, Opposites, and the Standard Progressive Matrices. Thus, the Gustafsson (1984) higher-order model was defined by a total of 20 indicators. As Raven's association with *g* was of central interest to this investigation, the correlation matrix reported by Gustafsson (1984) was first simulated in SPSS to yield observed data. Then, with the observed simulated data, the four tests which were split into two halves by Gustafsson (1984) were "recombined" to form four individual test composites. For example, the Raven odd items composite and the Raven even items composite variables in the originally simulated raw data were summed together to form an overall Raven total composite variable. The same procedure was applied to the Mental Folding, Card Rotation, and Opposites composite variables. Thus, in this investigation, the original 16 tests administered in Gustafsson (1984) were included as measures of intelligence, and the association between the SPM's total scores and *g* and a nested group-level factor could be estimated via the bifactor model. Based on Gustafsson's (1984) higher-order model, four of the 16 tests were classified as measures of g_v (Copying, Card Rotation, Hidden Patterns, Group Embedded Figures Test), four as measures of g_f (Standard Progressive Matrices, Mental Folding, Disguised Pictures, Disguised Words), two as measures of g_s (Number Series, Letter Grouping), two as measures of short-term memory (g_{sm}; Auditory Number Span, Auditory Letter Span), and four as measures of g_c (Opposites, Swedish Achievement, Mathematics Achievement, English Achievement).

5.3. Data analysis

A series of four bifactor models were tested in this investigation. First, as can be seen in Fig. 1 (Marshalek), the Marshalek et al. (1983) data were modeled as a bifactor model with one first-order general factor defined by 19 indicators and five nested group-level factors (g_f, g_s, g_q, g_c, g_{sm}) each defined by four indicators (with the exception of g_{sm} which was defined by three indicators). The Reading Comprehension and Reading Vocabulary indicator uniquenesses were allowed to covary to account for shared method variance. The MISTRA-CAB data were modeled as a bifactor model with one first-order general factor defined by 15 indicators and four nested group-level factors (g_f, g_s, g_{sm}, g_c; see Fig. 1, MISTRA-CAB). Two of the CAB subtests were not classifiable within a nested factor; however, they were included in the model to help further define the breadth of the *g* factor. The MISTRA-WAIS data were modeled as a bifactor model (see Fig. 1, MISTRA-WAIS) with on first-order general factor defined by 12 indicators and three nested group-level factors (g_f, Freedom from Distractibility, g_c). Finally, the Gustafsson (1984) data were modeled as a bifactor model (see Fig. 1, Gustafsson) with one first-order general factor defined by 16 indicators and four nested group-level factors (g_f, g_v, g_c, g_{sm}). There was an insufficient amount of covariance between the Number Series and Letter Grouping indicators to form a statistically significant nested factor, independent of the *g* factor. The Disguised Pictures and Disguised Words indicator uniquenesses were allowed to covary to account for shared method variance.

All models were estimated via maximum likelihood and all models were identified by constraining the latent variable variances to 1. A model was considered acceptably well-fitting in the event that RMSEA and SRMR values were approximately .08 or smaller and CFI and TLI values were approximately .95 or greater. Unfortunately, none of the papers from which the correlation matrices were obtained published the internal consistency reliabilities associated with any of the tests, including Raven's. Based on Raven, Raven, and Court (2000) and Raven, Raven, and Court (1998), the SPM and the APM test scores have been reported to be associated with internal consistency reliabilities of approximately .85. Consequently, for the purposes of estimating the amount of test specificity associated with the Raven's scores for any particular sample/model, an internal consistency reliability estimate of .85 associated was used.

6. Results

The first bifactor model (Marshalek) was observed to be acceptably well-fitting, $\chi^2(132) = 258.72$, $p < .001$, RMSEA = .063, SRMR = .047, TLI = .944, CFI = .957. As can be seen in Table 1, the APM was associated with a standardized *g* loading of .76 and secondary loading of .31 on the nested g_v factor. It will be noted that several subtests were associated with numerically more substantial *g* loadings than the APM (e.g., Arithmetic Concepts = .85; Arithmetic Applications = .79). Furthermore, the APM's secondary g_v loading was comparable in size to that observed for the other subtests which loaded on the nested g_v factor. Thus, the APM was not observed to be a remarkably high *g* loading test, nor a pure measure of *g* in this model.

The second bifactor model (MISTRA-CAB) was observed to be acceptably well-fitting, $\chi^2(77) = 226.84$, $p < .001$, RMSEA = .067, SRMR = .042, TLI = .933, CFI = .951. As can be seen in Table 2, the SPM was associated with a standardized *g* loading of .65 and secondary loading of .41 on the nested g_f factor. As per the Marshalek model, it will be noted that several subtests were associated with numerically

[2] The possibility of examining the MISTRA data across all three batteries in a single bifactor model was considered. However, it was effectively impossible to specify a well-fitting bifactor model in the absence of a large number of cross-loadings and correlated residuals.

[3] The third test battery included in the MISTRA data set, the Hawaii battery, was not a purposely developed intelligence battery. Instead, it was a collection of 15 subtests chosen to investigate intelligence in twins with a particular focus on spatial ability (Johnson & Bouchard, 2011). Based on my bifactor analyses (available upon request), there were no group-level factors defined by more than two subtests associated with the Hawaii battery (other than mental rotation). Consequently, as Raven's would probably be best considered a secondary indicator of g_f or g_v, the Hawaii battery was not selected for analysis in this investigation.

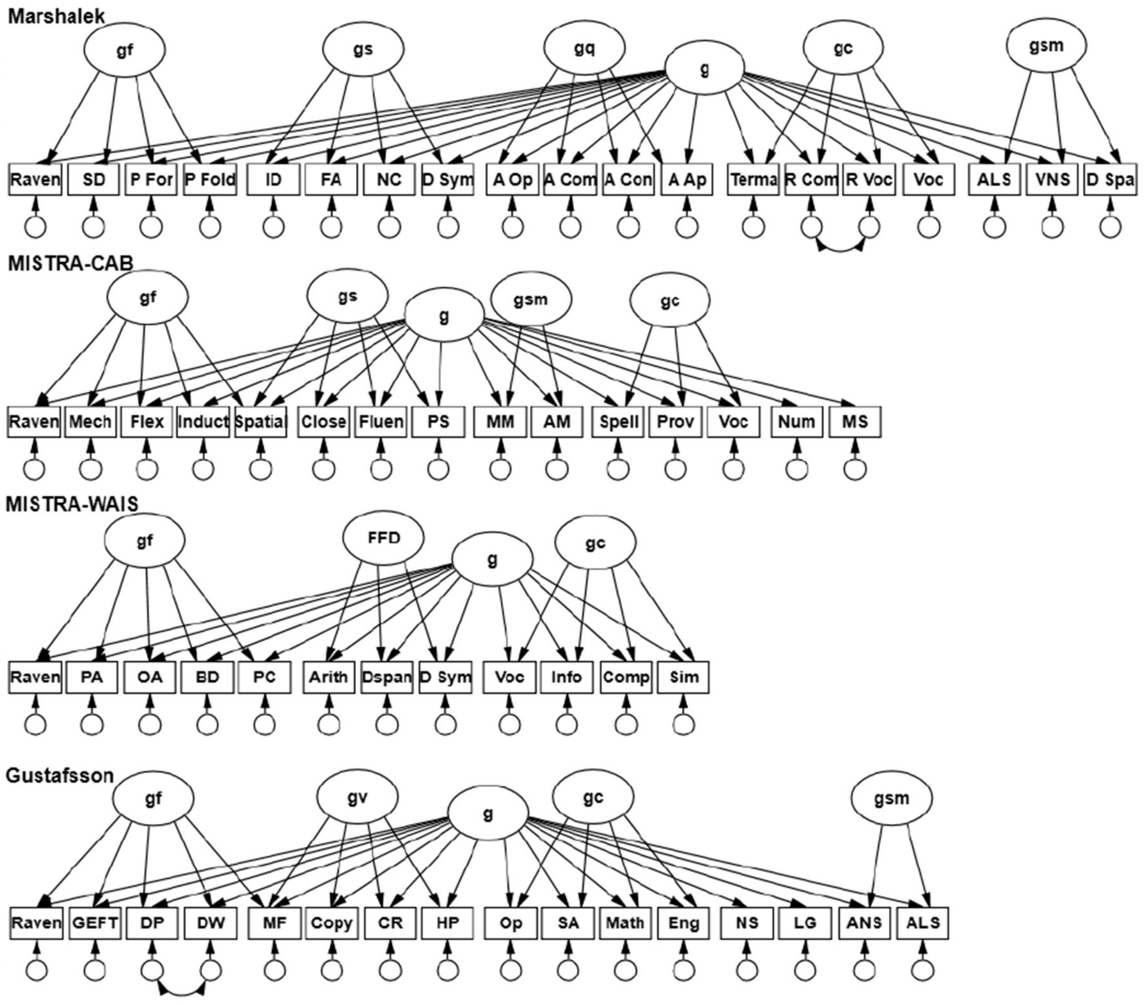

Fig. 1. The four bifactor models tested in this investigation; Marshalek = Marshalek et al. (1983) data (see Table 1 for full subtest names); MISTRA-CAB = MISTRA (Johnson & Bouchard, 2011) CAB data (see Table 2 for full subtest names); MISTRA-WAIS = MISTRA (Johnson & Bouchard, 2011) WAIS data (see Table 3 for full subtest names); and Gustafsson = Gustafsson (1984) data (see Table 4 for full subtest names).

more substantial *g* loadings than the SPM's (e.g., Numerical = .72; Fluency = .77). Furthermore, the SPM's secondary loading was comparable in size as the other subtests which loaded onto the nested g_f factor.

The third bifactor model (MISTRA-WAIS) was observed to be acceptably well-fitting, $\chi^2(42) = 98.95$, $p < .001$, RMSEA = .056, SRMR = .033, TLI = .963, CFI = .976. As can be seen in Table 3, the SPM was associated with a standardized *g* loading of .72 and secondary loading of .26 on the nested g_v factor. As per Models 1 and 2, it will be noted that several subtests were associated with numerically more

Table 1
Completely standardized bifactor model solution: Marshalek et al. (1983) data.

	g	g_v	g_s	g_q	g_c	g_sm
Advanced Progressive Matrices	.76	.31				
Surface Development	.61	.54				
Paper Formboard	.50	.50				
Paper Folding	.64	.54				
Identical Pictures	.44		.46			
Find As	.39		.46			
Number Comparison	.35		.63			
Digit Symbol	.41		.69			
Arithmetic Operations	.81			−.02		
Arithmetic Comprehension	.77			.46		
Arithmetic Concepts	.85			.33		
Arithmetic Applications	.79			.29		
Terman Concept Mastery	.77				.40	
Reading Comprehension	.73				.36	
Reading Vocabulary	.70				.54	
Vocabulary	.65				.65	
Auditory Letter Span	.39					.47
Visual Number Span	.50					.57
Digit Span	.44					.68

Note. $N = 241$; all loadings were statistically significant ($p < .05$); the Reading Comprehension and Reading Vocabulary uniquenesses were correlated at .36, $p < .001$.

Table 2
Completely standardized bifactor model solution: MISTRA (Johnson & Bouchard, 2011) data — CAB and Raven's.

	g	g_f	g_s	g_sm	g_c
Standard Progressive Matrices	.65	.41			
Mechanical	.39	.40			
Flexibility	.61	.16			
Inductive	.69	.20			
Spatial	.42	.42	.33		
Closure	.58		.12		
Fluency	.77		.01		
Perceptual Speed	.64		.66		
Meaningful Memory	.59			.39	
Associative Memory	.49			.39	
Spelling	.79				.19
Proverbs	.68				.36
Vocabulary	.75				.61
Numerical	.82				
Memory Span	.60				

Note. $N = 433$; all loadings were statistically significant ($p < .05$).

Table 3
Completely standardized bifactor model solution: MISTRA (Johnson & Bouchard, 2011) data — WAIS and Raven's.

	g	g_v	g_{sm}	g_c
Standard Progressive Matrices	.72	.26		
Picture Arrangement	.48	.27		
Object Assembly	.35	.66		
Block Design	.59	.44		
Picture Completion	.57	.34		
Arithmetic	.75		.18	
Digit Span	.48		.50	
Digit Symbol	.55		.28	
Vocabulary	.78			.48
Information	.79			.31
Comprehension	.65			.43
Similarities	.67			.30

Note. $N = 433$; all loadings were statistically significant ($p < .05$).

substantial g loadings than the SPM's (e.g., Information = .79; Arithmetic = .75). Furthermore, the SPM's secondary g_v loading was comparable in size as the other subtests which loaded onto the nested g_v factor.

Finally, the fourth bifactor model (Gustafsson) was observed to be acceptably well-fitting, $\chi^2(88) = 238.22$, $p < .001$, RMSEA = .042, SRMR = .029, TLI = .968, CFI = .977. As can be seen in Table 4, the SPM was associated with a standardized g loading of .58 and secondary loading of .26 on the nested g_f factor. Again, as per the first three models, it will be noted that several subtests were associated with numerically more substantial g loadings than the SPM's (e.g., Number Series = .79; Mathematics = .79). Furthermore, the SPM's secondary g_f loading was comparable in size to that observed for the other subtests on the nested g_f factor.

Across all four models, Raven's was associated with a mean g loading of .68. By comparison, the mean g loading across all subtests and all four models was comparable at .60. With respect to the nested g_f/g_v factors, Raven's mean loading was equal to .31, which was very comparable in magnitude to the mean g_f/g_v loading of .35 associated with all g_f/g_v subtests. Finally, if Raven's test scores are, on average, associated with approximately 85% true score variance, and Raven's is associated with a unique g loading of ≈ .70 and a unique g_f loading of ≈ .30, then Raven's true score variance may be partitioned approximately in the following manner: 50% g, 10% g_f, and 25% test specific.[4]

7. Discussion

Across all four models, Raven's was observed to load onto g at approximately .70 and onto a secondary g_f/g_v factor in much the same manner as other g_f/g_v type subtests (≈ .30). Furthermore, across all models, two or more subtests were observed to load more appreciably on g than Raven's, often substantially so. The results of this investigation accord very well with the bifactor results reported by Gignac (2008), where Raven's was observed to load onto g at .55 and onto a nested g_v factor at .36. Consequently, it may be suggested that Raven's is neither a pure measure of g, nor even a remarkably high g loading test, which is in contrast to the fairly commonly expressed view in the literature (Deary & Smith, 2004; Jensen, 1998; Kline, 2000).

In light of the results of this investigation, it may be specified that Raven's test scores are approximately 50% g related variance, 10% g_f related variance, and 25% test specificity. Many tests would be expected to exhibit such a psychometric profile. In fact, based on several bifactor model analyses of the Wechsler scales, Matrix Reasoning, a test very similar to Raven's, has been reported to be associated with g and secondary group-level loadings very similar to those reported here with respect to Raven's (Canivez, 2014; Gignac, 2006b; Gignac & Watkins, 2013; Watkins, 2010). Given that Matrix Reasoning scores tend to be associated with internal consistency reliability of approximately .90 (Wechsler, 2008), its level of test specificity would also be approximately 25–30%. By contrast, a subtest such as Vocabulary appears to be associated with less test specificity. Specifically, across the three models tested in this investigation in which the WAIS Vocabulary subtest was included, it was associated with a mean g loading of a .73 and a mean g_c loading of .58. Thus, as the Vocabulary subtest scores have been reported to be associated with an internal consistency reliability of .95 (Wechsler, 1955), its reliable variance may be partitioned in the following manner: 53% g, 34% g_c, and 3% test specificity.

Table 4
Completely standardized bifactor model solution: Gustafsson (1984) data.

	g	g_f	g_v	g_c	g_{sm}
Standard Progressive Matrices	.58	.26			
Group Embedded Figure	.55	.20	.34		
Disguised Pictures	.21	.28			
Disguised Words	.33	.04			
Mental Folding	.52	.49	.32		
Copying	.53		.56		
Card Rotations	.50		.40		
Hidden Patterns	.55		.48		
Opposites	.59			.47	
Swedish Achievement	.72			.63	
Mathematics	.79			.15	
English	.66			.49	
Number Series	.79				
Letter Grouping	.68				
Auditory Number Span	.27				.63
Auditory Letter Span	.35				.63

Note. $N = 981$; the Disguised Words .04 loading was not statistically significant ($p = .455$); the Disguised Pictures and Disguised Words uniquenesses were correlated at .34, $p < .001$.

The results reported in this investigation are not peculiar to the bifactor model. In fact, comparable results were observed based on corresponding higher-order models (full results available upon request). For example, with respect to the Marshalek et al. (1983) sample, the APM had an association of .66 with g and an association of .50 with the g_v first-order factor residual (i.e., based on a Schmid–Leiman transformation of the higher-order model solution). The bifactor model was chosen because the associations between the tests and the latent variables are estimated directly, rather than indirectly via Schmid–Leiman transformations in the higher-order model case (Gignac, 2007).

Theoretically, the results of this investigation suggest that the positioning of Raven's as the unique essence of general intellectual functioning is likely unjustified. For example, Marshalek et al. (1983) contended that Raven's was fundamental to general intelligence, because it is one of the most complex tasks to execute, as it draws substantially upon "… executive assembly and control processes that structure and analyse the problem, assemble a strategy of attack on it, monitor the performance process, and adapt these strategies as performance proceeds…" (p. 124). Although complexity may play a role in the degree to which a task is related to g, Raven's does not appear to be uniquely complex in that respect. Even based on Marshalek et al.'s multidimensional scaling analyses, a quantitative reasoning test (Necessary Arithmetic Operations) was observed to be just as central to g as Raven's. Correspondingly, if there was something noteworthy in the pattern of g loadings across the four models tested in this investigation, it was that a quantitative reasoning test was observed to be fairly consistently associated with the most substantial g loading. Additionally, within the WAIS-IV normative sample, Arithmetic has been observed to be the most substantial g loading subtest (Gignac & Watkins, 2013). Although widely considered a core dimension of g (Carroll, 1993), few, if any, researchers would claim quantitative reasoning to be the unique essence of g.

The results of this investigation may also have potential implications for interpretations of the Flynn effect (Flynn, 2012; Lynn, 1982). Specifically, based on the results of this investigation, Raven's test scores appear to be associated with approximately 25% true score (i.e., reliable)

[4] $S_g^2 = (.70^2) * 100 \approx 50\%$; $S_{gf}^2 = (.30^2) * 100 \approx 10\%$; $85\% - (50\% + 10\%) = 25\%$.

variance that is unique, i.e., not shared with either g or g_f. Thus, Flynn's (1984) "baffling" observation that the Flynn effect appears to be operating in a substantial manner on Raven's test scores, but there has been no corresponding systematic increases in SAT scores or the observation of a cultural renaissance (Flynn, 1987), may be because the Flynn effect is operating nearly entirely upon the non-negligible amount of test specific reliable variance associated with Raven's. Similarly, the reported increases in intelligence test scores as measured by Raven's via "cognitive training" (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) may also operate substantially, or entirely, upon Raven's 25% test specific reliable variance. In fact, in a recent multi-group longitudinal latent variable investigation which included the RAPM, it was found that test score increases due to cognitive practice were not observed to be due to a common factor (Estrada, Ferrer, Abad, Román, & Colom, 2015).

Why it is that some intelligence researchers promulgated the notion that Raven's was a relatively pure measure of g, even in the face of contradictory evidence, not to mention clearly inappropriate citations to support such a contention (i.e., Burke, 1958), is difficult to explain with any confidence. To some degree, the controversial "Spearman's hypothesis" of black-white differences in intelligence was analysed and interpreted within the context that Raven's was a pure measure of g (see, for example, Jensen, 1987).[5] Thus, given the notoriety associated with this area of research, it may be understood why the argument was repeated in the literature. Relatively superficial readings of the psychometric literature by many very likely helped, as well. Unfortunately, the notion that Raven's is a pure measure of g made its way into the broader field of psychology. Consequently, researchers occasionally interpret the failed observation of an association between an independent variable and RPM scores as a failure to observe an association between an independent variable and g (e.g., Zhu et al., 2010).

In comparison to administering Raven's, researchers have a much more attractive option to obtain an estimate of g, both with respect to estimating g more accurately, as well as administration time. The option involves the administration of a small number of subtests from a comprehensive intelligence battery. For example, researchers could administer the Similarities, Digit Span Backward, Matrix Reasoning, and Digit Symbol subtests from the Wechsler scales. Based on Axelrod's (2001) work with the Wechsler Adult Scale of Intelligence—III (Wechsler, 1997), such a subtest combination would take, on average, approximately 14 min to administer, in comparison to the approximate 40 min that it can take to administer Raven's (Arthur & Day, 1994). Additionally, total composite scores from the four-subtest combination described above would correlate with g at approximately .93,[6] which is meaningfully superior to Raven's $\approx$ .70 association with g. Several other subtest combinations with high validity have been reported for the Wechsler scales (Sattler & Ryan, 2009). Researchers would also have some capacity to explore possible effects that may be unique to several lower-order dimensions of intelligence (i.e., g_c, g_{sm}, g_f, and g_s). Finally, a meaningful, if far from perfect, g latent variable could be defined by the four subtests described above, as well as combinations described by others (e.g., Sattler & Ryan, 2009). The recommendation made here is consistent with Haier et al. (2009) who recommended broader measurement of cognitive abilities to estimate g in neuroscientific and intelligence research, in order to help increase the chances of obtaining consistent results across studies. Ultimately, a broad construct such as general intelligence would be expected to require scores obtained from a variety of test contents and test formats.

7.1. Limitations

In this investigation, the associations between the subtests and the latent variables were not disattenuated for imperfect reliability in the subtest scores. Such a procedure was not possible, as the subtest score reliabilities were not reported in any of the investigations. However, as this investigation was particularly focussed upon Raven's, a subtest with a large number of items (30+), and a previously reported test score internal consistency reliability of .85 (Raven et al., 2000), it is very doubtful that Raven's was in any way biased against.

Another limitation associated with this investigation is related to the quality of the group-level factors upon which Raven's was specified to load. Raven's is probably best considered a measure of g_f (Mackintosh, 2011). However, in none of the models was an especially impressive g_f latent variable defined, if only because there was an insufficient number of proper g_f tests from which to choose. Perhaps the fact that Raven's was observed to load meaningfully upon group-level factors that could be variously described as g_v, g_v/g_f, or g_f is testimony to just how non-remarkable a test it is. It will be noted that a number of other models were tested in this investigation with different subtest combinations. In no case where the g factor was defined by a relatively acceptable number of subtests (10+) did Raven's yield the highest-loading on g and a zero loading on the corresponding secondary group-level factor, when a group-level g_f/g_v factor appeared.

Relatedly, the implications of these results are principally relevant only to g. Whether Raven's may be considered a pure measure of g_f is a separate empirical question. As noted by Cattell (1980) and Jensen (1980b), and highlighted by Colom and Garćia-López (2002), Raven's consists exclusively of figural type items, which is a rather narrow representation of g_f. By contrast, the Culture Fair Intelligence Test (CFIT; Cattell & Cattell, 1960), which consists of four subtests of non-verbal fluid reasoning (series, analogies, matrices, classifications), may be regarded as a better indicator of g_f. Whether a latent variable derived from the four CFIT subtests is a near perfect representation of g_f remains to be determined. According to Lohman and Lakin (2011), g_f consists of sequential reasoning, quantitative reasoning, and inductive reasoning. The CFIT may be best described as a measure of inductive reasoning and to some degree sequential reasoning, however, it is not at all a measure of quantitative reasoning. Consequently, neither Raven's nor the CFIT are likely completely representative indicators of g_f.

Finally, although the sizes of the samples included in this investigation were adequate for the purposes of conducting the analyses, it should be emphasized that the representativeness of samples is a very significant consideration in the degree to which the results can be interpreted validly in an investigation such as this one (i.e., generalizable to the population). Unfortunately, a relatively small amount of details were published relevant to the Marshalek et al. (1983) and Gustafsson (1984) samples. However, as these two samples were based on high-school and primary school students, respectively, rather than a homogenous group such as university students, it may be reasonable to suppose that the samples were reasonably representative of their respective populations. The MISTRA sample was based on twins and members of the community; consequently, it should be reasonably representative, as well. Nonetheless, the validity of the results reported in this investigation should be acknowledged to be contingent upon the degree to which they were in fact representative.

8. Conclusion

The notion that a single test score may best represent any particular psychological construct, much less a construct as abstract as general intelligence, is arguably contrary to psychological measurement theory (Campbell & Fiske, 1959). If general intelligence is a hypothetical entity postulated to represent the phenomenon of the positive manifold, positive correlations between diverse cognitive ability tests, then it would seem unlikely that it could be measured with a single subtest score.

[5] Curiously, Jensen (1987) cited Agrawal, Sinha and Jensen (1984) to support the claim that Raven's is an even better indicator of g than a comprehensive battery such as the WAIS. However, Agrawal et al. administered only the Standard Progressive Matrices (Raven, 1960) in their investigation, which made any comparisons with the WAIS impossible.

[6] I estimated the value of .93 based on the WAIS-IV normative sample correlation matrix and the Tellegen and Briggs (1967) short-form validity formula.

Ultimately, Spearman (1927, 1946) only theorised that the unique essence of *g* was the eduction of relations and correlates. The accumulated empirical research relevant to working memory (see Conway & Kovacks, 2013) and processing speed (see Grudnik & Kranzler, 2001; Jensen, 2006) would suggest that Spearman was wrong on this matter.

Finally, it would be a disservice to the intelligence research community if this investigation was used as evidence to suggest that Raven's is a poor test of cognitive ability. The evidence suggests that it is a good test of *g* and g_f/g_v. Overall, there is likely nothing particularly noteworthy or special about Raven's and its relationship to *g* or any other particular factor. Instead, it is probably best recognised as one of several good quality tests that can be used to help define a *g* factor.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288–318.

Agrawal, N., Sinha, S. N., & Jensen, A. R. (1984). Effects of inbreeding on Raven matrices. *Behavior Genetics*, *14*(6), 579–585.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*(2), 394–403.

Ashton, M. C., Lee, K., & Vernon, P. A. (2001). Which is the real intelligence? A reply to Robinson (1999). *Personality and Individual Differences*, *30*, 1353–1359.

Axelrod, B. N. (2001). Administration duration for the Wechsler Adult Intelligence Scale—III and Wechsler Memory Scale—III. *Archives of Clinical Neuropsychology*, *16*(3), 293–301.

Basso, A., De Renzi, E., Faglioni, P., Scotti, G., & Spinnler, H. (1973). Neuropsychological evidence for the existence of cerebral areas critical to the performance of intelligence tasks. *Brain*, *96*(4), 715–728.

Brody, N. (1987). Jensen, Gottfredson, and the black–white difference in intelligence test scores. *Behavioral and Brain Sciences*, *10*(03), 507–508.

Brunner, M. (2008). No *g* in education? *Learning and Individual Differences*, *18*(2), 152–165.

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *The Journal of Genetic Psychology*, *93*(2), 199–228.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Canivez, G. L. (2014). Construct validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. *School Psychology Quarterly*, *29*(1), 38–51.

Canivez, G. L. (2015). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer, & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements*. Gottingen, Germany: Hogrefe Publishers (in press).

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

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York: NY: Pergamon.

Cattell, R. B. (1980). They talk of some strict testing of us — Pish. *Behavioral and Brain Sciences*, *3*, 336–337.

Cattell, R. B., & Cattell, A. K. S. (1960). *Culture fair intelligence test: Scale 2*. Champaign, IL: IPAT.

Cohen, J. (1957). The factorial structure of the WAIS between early adulthood and old age. *Journal of Consulting Psychology*, *21*(4), 283.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*(3), 445–451.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by *g*. *Intelligence*, *32*, 277–296.

Conway, A. R. A., & Kovacks, K. (2013). Individual differences in intelligence and working memory: A review of latent variable models. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 233–270). San Diego, CA: Academic Press.

Corben, L. A., Georgiou-Karistianis, N., Fahey, M. C., Storey, E., Churchyard, A., Horne, M., et al. (2006). Towards an understanding of cognitive function in Friedreich ataxia. *Brain Research Bulletin*, *70*(3), 197–202.

Crawford, J. R., Deary, I. J., Allan, K. M., & Gustafsson, J. E. (1998). Evaluating competing models of the relationship between inspection time and psychometric intelligence. *Intelligence*, *26*(1), 27–42.

Day, E. A., Arthur, W., Bell, S. T., Edwards, B. D., Bennett, W., Mendoza, J. L., et al. (2005). Ability-based pairing strategies in the team-based training of a complex skill: Does the intelligence of your training partner matter? *Intelligence*, *33*(1), 39–65.

Deary, I. J., & Smith, P. (2004). Intelligence research and assessment in the United Kingdom. In R. J. Sternberg (Ed.), *International handbook of intelligence* (pp. 1–48). Cambridge, UK: Cambridge University Press.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of Personality*, *73*(4), 825–858.

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, *50*, 93–99.

Eysenck, H. J. (1998). *Dimensions of personality*. New Brunswick, NJ: Transaction Publishers.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*(1), 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 countries: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191.

Flynn, J. R. (2012). *Are we getting smarter?: Rising IQ in the twenty-first century*. Cambridge University Press.

Gignac, G. E. (2006a). Evaluating subtest '*g*' saturation levels via the Single Trait-Correlated Uniqueness (STCU) SEM approach: Evidence in favour of crystallized subtests as the best indicators of '*g*'. *Intelligence*, *34*, 29–46.

Gignac, G. E. (2006b). The WAIS-III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences*, *27*(2), 73–86.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, *42*(1), 37–48.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: *g* as superordinate or breadth factor? *Psychology Science*, *50*(1), 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*(5), 639–662.

Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence*, *29*(6), 523–535.

Gustafsson, J. -E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407–434.

Haier, R. J., Colom, R., Schroeder, D. H., Condon, C. A., Tang, C., Eaves, E., et al. (2009). Gray matter and intelligence factors: Is there a neuro-g? *Intelligence*, *37*(2), 136–144.

Hakstian, A. R., & Cattell, R. B. (1975). *The comprehensive ability battery*. Champaign, IL: Institute for Personality and Ability Testing.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak, & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234–259). New York: Oxford University Press.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jensen, A. R. (1973). Level I and level II abilities in three ethnic groups. *American Educational Research Journal*, *10*(4), 263–276.

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

Jensen, A. R. (1980b). Author's response. Précis of bias in mental testing. *Behavioral and Brain Sciences*, *3*, 359–368.

Jensen, A. R. (1987). Further evidence for Spearman's hypothesis concerning black–white differences on psychometric tests. *Behavioral and Brain Sciences*, *10*(03), 512–519.

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

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam, Netherlands: Elsevier.

Johnson, W. (2012). How much can we boost IQ?: An updated look at Jensen's (1969) question and answer. In A. M. Slater, & P. C. Quinn (Eds.), *Developmental psychology: Revisiting the classic studies* (pp. 118–131). London: Sage Publications Ltd.

Johnson, W., & Bouchard, T. J. (2011). The MISTRA data: Forty-two mental ability tests in three batteries. *Intelligence*, *39*(2), 82–88.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, *32*(1), 95–107.

Kline, P. (2000). *The handbook of psychological testing*. London: Routlege.

Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric *g*: Unitary process or a number of independent processes? *Intelligence*, *15*(4), 397–422.

Llabre, M. M. (1984). Standard progressive matrices. In D. J. Keyser, & R. C. Sweetland (Eds.), *Test critiques*, *Vol. 1*. (pp. 595–602). Missouri: Test Corporation of America.

Lohman, D. F., & Lakin, J. M. (2011). Reasoning and intelligence. In R. J. Sternberg, & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 419–441) (2nd ed.). New York: Cambridge University Press.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. Dunnette, & L. M. Hough (Eds.), (2nd ed.). *Handbook of industrial and organizational psychology*, *Vol. 3*. (pp. 1–59). . Palo Alto, CA: Consulting Psychologists Press.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *297*, 222–223.

Mackintosh, N. J. (2011). *IQ and human intelligence*. Oxford: Oxford University Press.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107–127.

Martinez, M. E. (2013). *Future bright: A transforming vision of human intelligence*. New York, NY: Oxford University Press.

Meng, H. (2005) Hierarchical model of human intelligence factors on 37 ability tests. Unpublished manuscript.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2–20.

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*(5), 407–422.

Neisser, U. (1998). Introduction: Rising test scores and what they mean. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 3–22). Washington, DC: American Psychological Association.

Raven, J. C. (1940). Matrix tests. *Mental Health*, *1*, 10–18.

Raven, J. C. (1941). Standardization of progressive matrices. *The British Journal of Medical Psychology*, *19*, 137–150.

Raven, J. C. (1960). *Guide to the standard progressive matrices*. London: H. K. Lewis.

Raven, J. C. (1966). *Advanced progressive matrices*. New York: Psychological Corporation.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Raven's progressive matrices and vocabulary scales*. New York: Psychological Corporation.

Raven, J. C., Court, J. H., & Raven, J. (1994). *Advanced progressive matrices: Sets I and II*. Manual for Raven's progressive matrices and vocabulary scales. Oxford: England7 Oxford Psychologists Press.

Raven, J. C., Raven, J. C., & Court, J. H. (1962). *Coloured progressive matrices*. London: HK Lewis.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 4. Advanced progressive matrices*. Oxford: Oxford Psychologist Press.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Raven manual: Section 3. Standard progressive matrices*. Oxford: Oxford Psychologist Press.

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

Rogers, W. A., Fisk, A. D., & Hertzog, C. (1994). Do ability-performance relationships differentiate age and practice effects in visual search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 710–738.

Rushton, J. P., & Skuy, M. (2001). Performance on Raven's matrices by African and White university students in South Africa. *Intelligence*, *28*(4), 251–265.

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. San Antonio, TX: Psychological Corporation.

Schellenberg, E. G., & Moreno, S. (2010). Music lessons, pitch processing, and *g*. *Psychology of Music*, *38*(2), 209–221.

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*(1), 40–56.

Snow, R. E., Lohman, D. F., Marshalek, B., Yalow, E., & Webb, N. M. (1977). *Correlational analyses of reference aptitude constructs (NR154-376 ONR technical report no. 5)*. Stanford, CA: School of Education, Stanford University.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Spearman, C. (1946). Theory of general factor. *British Journal of Psychology. General Section*, *36*(3), 117–131.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, *31*, 499–506.

Thorndike, R. L. (1986). Historical and theoretical perspectives. In R. L. Thorndike, E. P. Hagen, & J. M. Sattler (Eds.), *The Standford–Binet intelligence scale (Fourth edition). Technical manual*. (pp. 1–7). Itasca, IL: Riverside.

Vernon, P. E. (1947). The variations of intelligence with occupation, age, and locality. *British Journal of Statistical Psychology*, *1*(1), 52–63.

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence*, *7*(1), 53–70.

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 287–314). New York, NY: Lawrence Erlbaum Associates.

Walker, S. Y., Pierre, R. B., Christie, C. D. C., & Chang, S. M. (2013). Neurocognitive function in HIV-positive children in a developing country. *International Journal of Infectious Diseases*, *17*(10), e862–e867.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children—Fourth Edition among a national sample of referred students. *Psychological Assessment*, *22*(4), 782–787.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York: The Psychology Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (Third Edition). San Antonio, TX: Pearson Assessment.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Pearson Assessment.

Wilhoit, B. E., & McCallum, R. S. (2003). Cross-battery assessment of nonverbal cognitive ability. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 63–78). New York: Plenum.

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., et al. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*(2), 137–142.