How Well Is Psychometric g Indexed by Global Composites? Evidence From Three Popular Intelligence Tests

Matthew R. Reynolds, Randy G. Floyd, and Christopher R. Niileksela

Online First Publication, August 12, 2013. doi: 10.1037/a0034102

CITATION
How Well Is Psychometric $g$ Indexed by Global Composites? Evidence From Three Popular Intelligence Tests

Matthew R. Reynolds
University of Kansas

Randy G. Floyd
University of Memphis

Christopher R. Niileksela
University of Kansas

Global composites (e.g., IQs) calculated in intelligence tests are interpreted as indexes of the general factor of intelligence, or psychometric $g$. It is therefore important to understand the proportion of variance in those global composites that is explained by $g$. In this study, we calculated this value, referred to as hierarchical omega, using large-scale, nationally representative norming sample data from 3 popular individually administered tests of intelligence for children and adolescents. We also calculated the proportion of variance explained in the global composites by $g$ and the group factors, referred to as omega total, or composite reliability, for comparison purposes. Within each battery, $g$ was measured equally well. Using total sample data, we found that 82%–83% of the total test score variance was explained by $g$. The group factors were also measured in the global composites, with both $g$ and group factors explaining 89%–91% of the total test score variance for the total samples. Global composites are primarily indexes of $g$, but the group factors, as a whole, also explain a meaningful amount of variance.

Keywords: intelligence, composite reliability, omega, psychometric $g$, IQ

Intelligence tests are among the most popular and useful of all measurement devices in psychology (Flanagan & Harrison, 2012; Kaufman, 2009). The factor structure of the subtest scores of the intelligence test relative to the measurement structure of the test is the focus of much factor-analytic research. Multiple latent sources of variance in those subtest scores are described in that research via factor loadings (e.g., Canivez & Watkins, 2010; Reynolds, Keith, Fine, Fisher, & Low, 2007). The general factor (psychometric $g$ or simply $g$) is often interpreted as one latent source of variance in all subtest scores, with all subtests having positive nonzero loadings on $g$ (Jensen, 1998).

A global composite (e.g., an IQ) calculated as the sum of subtest scores, and not a score from an individual subtest, is intended to index $g$. Therefore, although subtest loadings on the $g$ factor (i.e., $g$ loadings) provide estimates of how strongly $g$ is measured in the individual subtests, and consequently, these $g$ loadings provide hints about $g$ measurement in a global composite when they are averaged across subtests, it is not clear how well global composites measure $g$. Such estimates are important if these composites are interpreted as indexes of $g$. In this study, the proportion of variance in global composite scores that is accounted for by the $g$ factor was calculated for three popular individually administered intelligence tests designed for children and adolescents.

Psychometric intelligence is described in multifactored hierarchical models (Gustafsson, 1984). One such model, based on Cattell–Horn–Carroll theory (CHC), is a general framework that currently includes $g$ at the apex, about eight to 10 broad cognitive ability factors (e.g., Verbal Comprehension–Knowledge [Gc], Novel Reasoning [Gf], Visual–Spatial Ability [Gv], Short-Term Memory [Gsm], and Processing Speed [Gs]), and numerous narrow abilities (Schneider & McGrew, 2012). CHC theory is a work in progress, and in fact Carroll (1993) and Horn (Horn & Blankson, 2005) disagreed on evidence of a theory of $g$, but CHC theory is a useful framework for research and applied practice because it describes intelligence at three levels of generality, with scores from intelligence batteries often reflecting each of these levels (Keith & Reynolds, 2010). The global composite (IQ) is considered an indicator of $g$, broad composite scores as indicators of CHC broad abilities (otherwise known as group factors), and individual subtest scores as indicators of narrow abilities. In this study, the focus was on the indicator of $g$ within a test battery, the global composite.

More than 85 years ago, Spearman (1927, Appendix A) proposed an estimate that was supposed to represent the $g$ loading of the global composite (also see Jensen, 1998, p. 104). More recently, model-based estimates calculated from structural equation models have been proposed (e.g., McDonald, 1999). McDonald has referred to these estimates as coefficient omega. In terms of classical test theory, omega is both a reliability and validity index because it is the true score variance in the composite that is attributed to a common factor (Bollen, 1989; Gustafsson & Åberg-
Omega is a general estimate and may be used to determine the saturation of one factor (e.g., \( g \)) in a composite even when there are multiple factors (e.g., \( g \) and group factors) contributing to the composite, and it may also be used to calculate the saturation of all of the factors contributing to the composite. The former estimate has been referred to as hierarchical omega (\( \omega_h \); Zinbarg, Yovel, Revelle, & McDonald, 2006), whereas the latter has been subsequently referred to as omega total (\( \omega_T \); Revelle & Zinbarg, 2009). That is, \( \omega_h \) may be used to quantify the saturation of \( g \) in a global composite, whereas \( \omega_T \) may be used to quantify the saturation of all of the common factors in a global composite (e.g., \( g \) and group factors) and is also therefore described as composite reliability. The distinction is important because despite the multifactorial nature of intelligence, in practice, global composites are mostly interpreted as indexes of factorial nature of intelligence, in practice, global composites are mostly interpreted as indexes of \( g \), not of intelligence in general (\( g + \) group factors). In this study, we compared these two estimates because although global composites are intended to index \( g \), it is also important to understand how well they index all of the common factors.

**Purpose**

Because global composite scores are interpreted as indexing \( g \), a latent construct that has been subjected to years of study and has a wide range of predictive validity (Jensen, 1998), it is especially important to estimate how well global composite scores measure this construct. The primary purpose of this study was to estimate the \( g \) saturation in global composites from three popular individually administered intelligence tests designed for children and adolescents. This \( g \) saturation was indexed by \( \omega_h \). The \( g \) saturation (\( \omega_h \)) was also compared with the saturation of all of the factors by calculating \( \omega_T \), or composite reliability.

**Method**

**Participants**

The participants were children and adolescents included in the norming samples for three intelligence tests. These norming samples were representative of the United States (Elliott, 2007b; Kaufman & Kaufman, 2004; Wechsler, 2003). The total samples for each test ranged from 2175 to 2200. There were 200 participants per 1-year age group for the second edition of the Differential Abilities Scales (DAS–II; Elliott, 2007a), 125 to 200 participants per 1-year age group for the second edition of the Kaufman Assessment Battery for Children (KABC–II; Kaufman & Kaufman, 2004), and 200 participants per 1-year age group for the fourth edition of the Wechsler Intelligence Scale for Children (WISC–IV; Wechsler, 2003).

**Measurement Instruments**

**Differential Abilities Scales–Second Edition (DAS–II).** The DAS–II (Elliott, 2007a) is an individually administered intelligence test developed for children and adolescents. Fifteen of the school-age battery subtests for children ages 7–17 were used in this study. The DAS–II provides composites representing six group factors: Nonverbal Reasoning, Verbal, Spatial, Verbal Short-Term Memory, Visual–Verbal Memory, and Cognitive Speed. The \( g \) factor is operationalized in the test as General Conceptual Ability (GCA), comprising six subtests. Composite score reliability estimates for the GCAs ranged from .96 to .97 in the ages used in this study (Elliott, 2007b, pp. 127–128).

**Kaufman Assessment Battery for Children–Second Edition (KABC–II).** The KABC–II is an individually administered measure of cognitive abilities developed for children and adolescents. Data from 16 subtests used with children ages 7–18 were included in this study. The KABC–II provides composites representing five group factors: Crystallized Knowledge, Fluid Reasoning, Visual–Spatial Ability, Long-Term Retrieval, and Short-Term Memory. The \( g \) factor is operationalized in the test as the Fluid–Crystallized Index (FCI), comprising 10 subtests. Composite score reliability estimates for the FCIs ranged from .96 to .97 in the ages used in this study (Kaufman & Kaufman, 2004, p. 88).

**Wechsler Intelligence Scale for Children–Fourth Edition (WISC–IV).** The WISC–IV is an individually administered measure of cognitive abilities developed for children and adolescents ages 6–16. Fifteen subtests were used in this study. The WISC–IV provides composites representing four group factors: Perceptual Reasoning, Verbal Comprehension, Processing Speed, and Working Memory. The \( g \) factor is operationalized in the test as the Full Scale IQ (FSIQ), comprising 10 subtests. Composite score reliability estimates for the FSIQs ranged from .96 to .97 in the ages used in this study (Wechsler, 2003, p. 35).

**Analytic Approach**

**Nested factor models.** Covariance matrices for total samples and for each 1-year age group for each test (DAS–II, KABC–II, WISC–IV) were created based on the subtest correlations and standard deviations for the norm samples provided in the test manuals (standard deviations for the KABC–II had to be obtained from raw data files that were available because they are not provided in the test manual). These matrices were used as input data. Mplus (Muthén & Muthén, 1998–2010) was used for all analyses.

Nested factor models were estimated for each intelligence test with the total sample and with each 1-year age group data (see Gustafsson & Balke, 1993; Reynolds & Keith, 2013). In a nested factor model, the \( g \) factor is modeled as a first-order factor, with direct effects on every subtest. More specific first-order factors (or group factors) are modeled for a subset of interrelated subtests; thus, there are direct effects from two latent factors (i.e., \( g \) and group) on most subtests. All of these factors are independent from one another. We used nested factor models, rather than higher order models, to account for the hierarchical and multidimensional nature of intelligence. Although higher order models, in which \( g \) has direct effects on first-order factors and indirect effects on subtests through the first-order factors, are more consistent with intelligence theory (Jensen, 1998), the nested factor models are more consistent with the scoring structure of intelligence tests (Keith & Reynolds, 2012). Regardless, nested factor models and higher order models typically produce similar \( \omega_h \) values because the average \( g \) loadings from the higher order model (calculated as total \( g \) effects, normally only indirect effects through the first-order factors) are similar to the average \( g \) loadings in the nested factor model (Zinbarg et al., 2006; see Niileksela, Reynolds, & Kaufman, 2013, for an example).
All subtests included in each test battery were included in the nested factor models so that the best estimates of g loadings were obtained, and because in many instances, it allowed for the group factors to be empirically identified without having to include equality constraints on the loadings. Group factors are often indexed by composites comprising two subtests. A factor not correlated with other factors and with only two subtest indicators is empirically underidentified. In these circumstances in which there are only two indicators, constraints must be included in the model (e.g., fix the two factor loadings to be equal) for proper model identification. However, test batteries often include additional measures of these group factors; therefore, this problem of underidentification was mostly circumvented in this study by including these additional subtests in the nested factor models. The estimates associated with these additional subtests, however, were not included the calculations of the omegas; only those associated with the global composite were included in those calculations.1

The DAS–II nested factor model included 13 subtests that loaded directly on the first-order g factor. In children ages 12 and younger, g was also indicated by Phonological Processing, although this subtest was not associated with an additional first-order factor. In addition to g, six first-order group factors associated with subsets of those subtests (i.e., two to three subtests) were included. These group factors were independent of each other and of g. Only one group factor was indicated by more than two subtests, so equality constraints were applied to the loadings of group factors with two indicators to properly identify the model. For purposes of calculating the omegas, however, only factor loadings and residual variances from the six subtests contributing to the GCA were used: Pattern Construction, Recall of Designs, Matrices, Sequential and Quantitative Reasoning, Word Definitions, and Verbal Similarities.

The KABC–II model included 16 subtests with direct loadings on g. The model included four first-order group factors because there was not reliable Fluid Reasoning variance remaining once g was accounted for (Gustafsson, 1984; Reynolds & Keith, 2007). Factor loadings and residual variances from 10 subtests contributing to the FCI were used in the omega calculations; the Gestalt Closure, Expressive Vocabulary, Hand Movements, Atlantis Delayed, and Rebus Delayed subtests were omitted.

The WISC–IV model included 15 subtests loading directly on g. There were four first-order group factors. Factor loadings and residual variances from 10 subtests contributing to the FSIQ were used in the omega calculations; the Picture Completion, Arithmetic, Information, Word Reasoning, and Cancellation subtests were omitted.

Model evaluation. We evaluated the models globally by examining the fit indexes. Global indexes of stand-alone fit included chi-square test of fit statistic, root-mean-square error of approximation (RMSEA), comparative fit index (CFI), and standardized root-mean-square residual (SRMR). Models with “good” global fit typically have RMSEA values close to or less than .05, CFI values close to or greater than .95, and SRMR values close to or less than .05 (Schermelleh-Engel, Moosbrugger, & Müller, 2003). In addition, the quality of the parameter estimates, as indicated by implausible values, was evaluated. Omega calculations were based only on models that had acceptable global fit and parameter estimates.

Omega calculations. Both $\omega_g$ and $\omega_T$ values were calculated from the nested factor model-based estimates. The estimates were calculated in the total sample and in each 1-year age group for each test. First, $\omega_g$ was calculated as the square of the sum of subtest g loadings divided by the total variance in the subtest scores included in the composite (Gustafsson, 2002). This estimate represented the proportion of variance in the global composite that was accounted for by g. The square root of $\omega_g$ represented the correlation between the composite score and latent g, which is the g loading for the composite (McDonald, 1999). Second, $\omega_T$ estimates were calculated; these estimates were based on all of the common factors in the nested factor model, including g and the other first-order group factors. The sum of subtest g loadings squared, and the sum of each group’s factor subtest loadings squared, were summed and divided by the total variance in the subtest scores.

Results and Discussion

Models for each test using the total samples and their respective fit indexes are shown in Figures 1, 2, and 3. Only the fit indexes for the total samples are reported for each test because the fit statistics for the age-specific models were similar (except that the model chi-square value was larger in the total samples because of their large sample sizes). All of the models fit very well. Some slight adjustments, noted in Table 1, were made to models for convergence purposes in some age groups. The bolded g loadings in Figures 1–3 were the estimates included in $\omega_g$ calculations; the bolded g and group factor loadings were included in the $\omega_T$ calculations.

As shown in Table 1, the global composites from each test—the GCA (Elliott, 2007a), the FCI (Kaufman & Kaufman, 2004), and the FSIQ (Wechsler, 2003)—measure their respective g factors similarly. In the total sample for the DAS–II, the g saturation ($\omega_g$) was .83, and across age groups, values ranged from .79 for 17-year-olds to .87 for 8- and 9-year-olds ($SD = .03$). In the total sample for the KABC–II, the g saturation was .82, and values ranged from .78 for 7- and 16-year-olds to .87 for 15-year-olds ($SD = .03$). Last, in the total sample for the WISC–IV, the g saturation was .83, and values ranged from .79 for 6-year-olds to .85 for 15-year-olds ($SD = .02$). Therefore, using total sample data within each test, 82%–83% of the variance in the global composite score was explained by g. About 82%–84% of the variance in the total composite scores within each test was explained by $g$ when the estimates were averaged across the age groups (see last row in Table 1).

As shown in the third column for each test in Table 1, although g was predominantly measured by the global composites, the group factors were also measured. In the total sample for the DAS–II, the combined g and group factor saturation ($\omega_T$) was .89, and values ranged from .87 for 17-year-olds to .93 for 8- and 9-year-olds ($SD = .02$) across age groups. In the total sample for

---

1 Models were estimated using only the subtests included in the global composites (i.e., other subtests were not included), and the findings were similar. Moreover, models were estimated that calculated omega values using all available subtests and creating a composite from those subtests. As per the principle of aggregation, the omega values were larger when all subtests were used to create a composite score. Interested readers should contact the first author to obtain those results.
the KABC–II, the combined $g$ and group factor saturation was .89, and values ranged from .87 for 8-year-olds to .92 for 15-year-olds ($SD = .02$). Last, in the total sample for the WISC–IV, the combined $g$ and group factor saturation was .91, and values ranged from .88 for 6- and 7-year-olds to .93 for 15-year-olds ($SD = .02$). Therefore, using total sample data within each test, 89%–91% of the variance explained in the global composite was explained by $g$ and the group factors. About 90%–91% of the global composite variance was explained by both $g$ and the group factors when the estimates were averaged across the age ranges (see last row in Table 1). The variance explained by the group-factors was much less than $g$, but the amount was not trivial, with an additional 5%–12% of the variance explained by these group factors. Global composites are primarily indexes of $g$, but they are also indexes of intelligence in general.

One practical implication of these findings is related to the calculation of confidence intervals around global composites. A simple “true-score” confidence interval for a global composite score, using the reliability estimate based on internal consistency, will be narrower than the construct-, or validity-, based confidence interval for a global composite as an estimate of $g$. If clinicians interpret global composites as estimates of $g$, then it seems to make more sense to use a validity-based confidence interval. For example, confidence intervals are often calculated from Cronbach’s alpha, which for global composites from the tests used in this study ranged from .96–.97. Using these estimates, along with the typical 15-point standard deviation, an average standard error of measurement could be calculated as $15\sqrt{1-.97}$, which equals 2.59. A 95% confidence interval approximately equals $\pm 5.00$ (i.e., $1.96 \times 2.59$) and for a score of 100 would range from 95 to 105. Alternatively, the validity-based confidence interval for the global composite as an estimate of $g$ is based on $\omega_h$ (i.e., .82–.84 using the total samples, or .78–.87 in various age groups). Therefore, an average standard error of measurement, for example, is equal to $15\sqrt{1-.83}$, or 6.18. A 95% confidence interval around the global composite as an estimate of $g$ is approximately ±12, and for a score of 100 would range from 88 to 112. The confidence interval

Figure 1. Differential Abilities Scales–Second Edition nested-factor model using the total sample data. Bolded $g$ and bolded $g$ and group factor estimates were used in the calculations of hierarchical omega and omega total, respectively. Residuals variances have been removed from the figure. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.
based on the construct of interest, therefore, is much larger when the valid variance related to \( g \) is considered.\(^2\) Hence, a reliability estimate based on internal consistency should not be confused with a validity estimate (see Schneider, 2013).

Our findings may be generalized to another study in which similar estimates were calculated with a different method. As noted, \( \omega_h \) is the saturation of \( g \) in a global composite, or said differently, it is the \( g \) loading squared. In a study by Keith, Kranzler, and Flanagan (2001), children completed two intelligence tests, the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001) and Cognitive Assessment System (CAS; Naglieri & Das, 1997). Second-order \( g \) factors from those batteries were found to be perfectly correlated. Thus, the authors also correlated the \( g \) factor from the WJ III with the global composite from the CAS. The resulting \( g \) loading—the correlation between \( g \) and global composite—was .79, which is a \( g \) saturation of 62%, and less than all of the \( \omega_h \) values in this study.

The CAS subtests are not as \( g \) saturated as subtests in the tests we included in this study (Canivez, 2011), which may account for this difference. Nevertheless, although \( \omega_h \) may seem abstract, it is conceptualized as the square of the \( g \) loading of a global composite.

One area for future research is related to \( g \) measurement as a function of \( g \). Recent research, including research with intelligence tests used in this study, has shown that \( g \) is measured less well in intelligence tests as \( g \) increases (Detterman & Daniel, 1989; Reynolds, 2013; Reynolds, Keith, & Beretvas, 2010; Tucker-Drob, 2010). For example, in one study using estimates from a one-factor \( g \) model of DAS–II group factor-based composites, omegas were shown to decrease substantially across the IQ range (Reynolds, 2013). In that study, the omega values, calculated from models

\(^2\) We thank a reviewer for suggesting that we report this information.
with DAS–II group factor based composites, at IQs of 100 were similar if not identical to values calculated from the models with the DAS–II subtests in this study. Thus, the findings may generalize; however, future research is needed for a definitive answer. If they do generalize, the implications are that the confidence intervals about global composites, used as indicators of $g$, become wider as $g$ increases (and “confidence” in $g$ measurement decreases).

## Limitations

We did not investigate whether $g$ exists or whether $g$ is a meaningful psychological construct. Rather, we investigated how much of $g$, if such a construct exists, is measured in global composites of three popular intelligence tests. In general, this approach regards $g$ as a reflective latent variable and not as a formative composite or an aggregate of multiple abilities. It is important to recognize, however, that the assumption of whether $g$ is a reflective construct is debatable (van der Maas et al., 2006). All models require assumptions, and some researchers may not recognize the existence of $g$ as a valid assumption for a model. Thus, it should be noted that omega values may also be calculated for each of the group factor indexes, without the assumption of $g$ or with $g$ partialled. Such decisions should be based on the theoretical model of the researcher. Estimating all of these values was beyond the purpose of this study.

Similarly, it should be noted that the intelligence tests used in this study were either fully or to some extent based on theory, and they most notably have been studied with CHC theory. We modeled the data, however, so that they were more closely aligned with the intended factor structure of the tests. We could have used CHC theory-based predictions to improve model fit, and in some cases, those were used when the models did not converge. Nevertheless, the focus of this study was not on the group factors but the $g$ factor. Future studies may investigate how our omission of theory influenced the results (including the use of a more theory-based $g$ factor model, the higher order factor model, rather than the nested factor model used in this study).

Figure 3. Wechsler Intelligence Scale for Children–Fourth Edition nested-factor model using the total sample data. Bolded $g$ and bolded $g$ and group factor estimates were used in the calculations of hierarchical omega and omega total, respectively. Residuals variances have been removed from the figure. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
Conclusion


Received April 8, 2013
Revision received July 8, 2013
Accepted July 9, 2013