Psychological Test and Assessment Modeling, Volume 60, 2018 (4), 393-401

Reliability and interpretation of total scores from multidimensional cognitive measures – evaluating the GIK 4-6 using bifactor analysis

Tobias Debatin[1], *Abdullah Aljughaiman*[2], *Mariam AlGhawi*[3],
Heidrun Stoeger[4], *Albert Ziegler*[1]

Abstract

It is often not trivial to interpret total scores from test batteries of cognitive ability, as the underlying set of items or subscales is typically not unidimensional. Additionally, in such cases, the reliability is not accurately estimated by coefficient alpha. The rarely addressed problem and possible solutions via bifactor analysis are presented without mathematical details. Applying this technique, factor structure and model-based reliability of the cognitive ability part of the Gifted Identification Kit 4–6 (GIK 4–6), a newly developed test battery to identify gifted students in the United Arab Emirates, were evaluated using confirmatory bifactor analyses. Results revealed that the total score is reliable (coefficient omega = .89) and primarily measures a general intelligence factor (*g*). The total score should best be interpreted as a blend of *g*, a reading factor, and a mathematics and science factor.

Keywords: reliability, bifactor analysis, cognitive ability, validity, unidimensionality

[1] *Correspondence concerning this article should be addressed to:* Tobias Debatin, Department of Educational Psychology and Research on Excellence, Friedrich-Alexander University Erlangen-Nuremberg, Regensburger Straße 160, 90478 Nuremberg, Germany; email: tobias.debatin@fau.de

[2] King Faisal University

[3] Hamdan bin Rashid Al Maktoum Foundation for Distinguished Academic Performance

[4] University of Regensburg

In research, latent variable modeling and item response modeling are often used to assess psychological constructs. Nonetheless, especially in practice, when applying a test battery or questionnaire it is common to assess a construct of interest by summing all item scores or subscale scores to form a total score. However, the appropriateness of this procedure depends on several factors. Focusing on the domain of cognitive ability tests, this article will outline how to interpret total scores and assess their reliabilities when the underlying set of items is not unidimensional. The approach based on bifactor analysis (Green & Yang, 2015; Reise, 2012) will then be applied to evaluate the cognitive ability section of the Gifted Identification Kit 4–6 (GIK 4–6; Ziegler & Stoeger, 2016), a newly developed test battery of the Hamdan bin Rashid Al Maktoum Foundation for Distinguished Academic Performance to identify gifted students in the United Arab Emirates. In addition to evaluating factor structure and reliability of the GIK 4–6, the article has two other goals. First, we would like to clarify reliability estimation and interpretation of total scores from multidimensional measures without using formulas or mathematical details. Second, we want to highlight why the problem is ubiquitous in cognitive ability testing and why bifactor analysis is particularly well suited to address the problem for cognitive ability tests.

Currently, the Cattell-Horn-Carroll theory (CHC theory) is the most widely accepted model describing the structure of human cognitive abilities (McGrew, 2009). The model is based on hundreds of factor analyses of cognitive ability tests. Several correlated group factors were found to characterize the broad scope of human cognitive abilities. Usually, a second-order *g* is also postulated, as the group factors are correlated. Accordingly, cognitive ability test batteries typically consist of different subscales that either correspond to group factors of the CHC theory or are a blend of some of these factors. The subscales are normally summed to a total score; therefore, assuming correlated subscales, a certain percentage of reliable variance of the total score should be due to group factors and another due to *g* (Brunner & Süß, 2005). When using cognitive ability test batteries, the focus is often on the total score, which is then transformed into an IQ score. This raises two problems: First, how to calculate reliability for such a total score and second, how to interpret the total score?

The most common, but not well understood by many users, reliability coefficient is coefficient alpha (Cortina, 1993; Dunn, Baguley, & Brunsden, 2014; Slocum-Gori & Zumbo, 2011). For coefficient alpha to properly assess reliability essential tau equivalence has to be given, which includes the assumption of unidimensionality (Green & Yang, 2015; Slocum-Gori & Zumbo, 2011). Even though unidimensionality is normally clearly not given in cognitive test batteries as discussed before, coefficient alpha is commonly reported. A much better option to assess reliability when unidimensionality is not given is coefficient omega (McDonald, 1999). Using a factor-analytic framework coefficient omega estimates the proportion of the true score variance of the total score to the total score variance. Depending on which model fits the data best, the true score variance estimate is based either on a single factor or on several factors. It is the proportion of variance of the total score due to all reliable factors (Green & Yang, 2015).

However, the proportion of reliable variance of a total score alone does not inform us as to how it should be interpreted. Although the set of items or subtests underlying a total score of a cognitive ability test is rarely unidimensional, this is not automatically a problem for a meaningful interpretation. For an easy interpretation, it is sufficient that a total score primarily measures the target construct (Reise, Moore, & Haviland, 2010). A set of items fulfilling this condition is called essentially unidimensional (e.g., Slocum-Gori & Zumbo, 2011). There are several methods to estimate the degree of unidimensionality. Davenport, Davison Liou, and Love (2015) advocated the proportion of variance accounted for by the first principle component as an indicator, and Ten Berge and Sočan (2004) proposed the explained common variance (ECV) as an index, which is the percentage of common variance explained by the first common factor. There are good reasons to use these methods in certain circumstances, but when it comes to interpreting total scores, omega hierarchical (ω_H; e.g., Zinbarg, Revelle, Yovel, & Li, 2005) based on bifactor analysis, sometimes also called direct hierarchical model, seems to be the best option available (Green & Yang, 2015; Reise, 2012). Omega hierarchical assesses the proportion of variance of a total score due to the general factor that all items have in common. The judgment that ω_H based on bifactor analysis is the preferred method to assess the degree of unidimensionality assumes that a bifactor model fits the data well and corresponds to theory. This is why the method seems especially suitable to evaluate cognitive ability tests. In recent years, it has been shown repeatedly that a bifactor model suits data from cognitive ability tests very well and additionally corresponds nicely to CHC theory (Morgan, Hodge, Wells, & Watkins, 2015; Murray & Johnson, 2013).

A bifactor model assumes a general factor that directly influences scores on all items and several less general group factors that only influence the scores on some items. For example, the items of a verbal intelligence subtest could be influenced by *g* and a verbal factor but not by a factor representing numeric skills. All factors in a bifactor model are uncorrelated. Concerning the representation of human cognitive abilities, it competes with higher-order models that assume correlated group factors and a second-order general factor, which results from a second factor analysis of the group factors (Beaujean, 2015). The main difference is that in higher-order models the influence of the general factor on the items is indirect through the group factors, whereas in bifactor models, a direct influence is assumed. Thereby, bifactor models emphasize the general factor, whereas higher-order models prioritize group factors.

Remember that coefficient omega calculates the proportion of variance of the total score due to all reliable factors to estimate the amount of true score variance of a scale. The advantage of uncorrelated factors as in the bifactor model is that the variance due to all reliable factors can easily be calculated by simply summing the variances of all factors. Thereby, bifactor analysis can be used to calculate coefficient omega as well as ω_H (Green & Yang, 2015). Additionally, how much each factor of the model contributes to the variance of the total score can be directly compared. Not only ω_H, which assesses the contribution of the general factor, but also omega coefficients for the group factors, which quantify the amount of variance each group factor contributes, can be computed. In the case where ω_H is rather low, making an unidimensional interpretation of the total score questionable, the omega coefficients of the group factors can be very helpful in interpreting what the total score measures and to which degree.

Two previous bifactor analyses of cognitive test batteries revealed very different saturations with *g* assessed by ω_H. Gignac and Watkins (2013) found ω_H of the total score from the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV; Wechsler, 2008) to range from .84 to .88, depending on the age group. Brunner and Süß (2005) found ω_H of the total score from the Berlin Intelligence Structure test (BIS test; Jäger, Süß, & Beauducel, 1997; for an English description, see Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002) to be .68. While the analysis of the WAIS-IV clearly justifies an interpretation of the total score as measuring mainly g, the total score of the BIS should rather be interpreted as a blend of *g* and several group factors. However, despite having a lower ω_H than the WAIS-IV, the saturation with *g* in the BIS is still high and most of the reliable total score variance is due to *g*.

Present study

The current study assessed the factor structure and omega coefficients of the cognitive ability section of the GIK 4–6 consisting of five subtests: Verbal Ability I, Preknowledge in Science, Verbal Ability II, Preknowledge in Mathematics, and Nonverbal Ability (Ziegler & Stoeger, 2016). Following is a short description of the five subtests.

In Test 1: Verbal Ability I, students are presented with short sentences. Each sentence contains a word pair in parentheses. Working as quickly as possible, students decide which of these two words completes each sentence in a meaningful way by crossing out the incorrect word. In Test 2: Preknowledge in Science, students are presented with a series of words in which the first letter of each word has been moved to the end of the word. Working as quickly as possible, students judge whether the word is related to a given topic from the United Arab Emirates science curriculum. In Test 3: Verbal Ability II, students read two texts that have been specifically developed for the United Arab Emirates. For each text, students answer a series of questions in a multiple true–false format. In Test 4: Preknowledge in Mathematics, students are presented with four pages containing 16 circles each. Each circle contains an arithmetic problem. Working as quickly as possible, students connect the circles in the order of increasing results. Once they have completed a page, they immediately start with the next one. In Test 5: Nonverbal Ability, students are presented with a series of composite figures. Each figure contains three rows of figural elements. The progression of the figural elements depicted across each row follows a certain construction rule. Students identify the construction rule by examining the first two rows and then apply the rule to correctly complete the final row (by choosing one out of four options).

According to the application manual of the GIK 4–6, Verbal Ability I and II are combined to assess verbal ability, Preknowledge in Science and Preknowledge in Mathematics are combined to assess preknowledge in science and mathematics and Nonverbal Ability stands alone. Based on this, we tested the following factor structure with confirmatory bifactor analyses: We assumed a *g* factor to influence all five subtests directly, a reading factor to influence Verbal Ability I and II, and a science and mathematics factor to influence Preknowledge in Science and Preknowledge in Mathematics. Similar to previous bifactor analyses of cognitive test batteries, we assumed the majority of the total score variance to be attributable to *g*, meaning an ω_H of higher than .6.

Method

Sample

Our sample consisted of 1,142 students from grades four to six from the United Arab Emirates. The mean age of the sample was 9.74 years (*SD* = 0.95) and the percentage of females was 59%.

Data analysis

Confirmatory bifactor analyses were performed to analyze factor structure and reliability of the cognitive ability section of the GIK 4–6. The reliability coefficients of the five subtests Verbal Ability I, Preknowledge in Science, Verbal Ability II, Preknowledge in Mathematics, and Nonverbal Ability ranged from .74 to .90.

We started the analysis by testing the following factor structure: A *g* factor was assumed to influence all five subtests directly, a reading factor to influence Verbal Ability I and II, and a science and mathematics factor to influence Preknowledge in Science and Preknowledge in Mathematics. All factors were constrained to be uncorrelated and the two loadings of each group factor were constrained to be equal, as recommended for latent factors with only two indicators. Variances of all factors were set to 1. The maximum likelihood (ML) estimator was used for all analyses. All analyses were calculated using M*plus* 6.0.04 (Muthén & Muthén, 1998-2010). Missing values were handled by using the full information maximum likelihood method. Next to the chi-square value, model fit was assessed following the criteria of Hu and Bentler (1999). Accordingly, a value close to .95 for the Comparative Fit Index (CFI), a value close to .06 for the root mean square error of approximation (RMSEA), and a value close to .08 for the standardized root mean squared residual (SRMR) were the cutoff criteria for good model fit.

Results

The initial model showed good CFI and SRMR values but a highly significant chi-square value, and the value concerning RMSEA was far away from our cutoff criterion for good model fit (see Table 1). There was also a problem with the assumed reading factor as the loadings to its indicators were estimated to be 0. Modification indices, especially the standardized expected parameter change (SEPC), indicated influential covariance between Verbal Ability I and Preknowledge in Science, which was not captured by the model. Verbal Ability I is designed to assess mainly reading speed, whereas the main intention of Preknowledge in Science is to assess familiarity with concepts of the science curriculum. Nonetheless, it seems very plausible that a strong reading speed component is also present in Preknowledge in Science, as both tests are severely time-limited and language-based. Accordingly, we decided to change the model by adding Preknowledge in Science as a third indicator of the reading factor. Again, a *g* factor was assumed to influence all five subtests directly; the reading factor was now specified to influence Verbal Ability I and II as well as Preknowledge in Science, and a science and mathematics factor to influence Preknowledge in Science and Preknowledge in Mathematics. All factors were constrained to be uncorrelated and we constrained the two loadings of the science and mathematics factor to be equal. Variances of all factors were set to 1. This model fit the data extremely well, as indicated uniformly by all model fit values (see Table 1) and fit the data significantly better than the first model ($\chi^2_{\text{diff}}(2) = 42.96$, $p < .001$). All standardized loadings can be found in Table 2.

Concerning the factor structure, these results provide strong evidence for a *g* factor and two reliable group factors, interpreted as a reading factor and a mathematics and science factor. This final model was used to calculate the omega coefficients in the next step.

Coefficient omega, which is the proportion of variance of the total score due to all reliable factors, was .89, 95% CI [.87, .91]. The proportion of variance of the total score due to the *g* factor, ω_H, was .64, 95% CI [.58, .69]. The proportion of variance of the total score due to the reading factor was .20, 95% CI [.15, .26] and for the mathematics and science factor it was .06, 95% CI [.05, .06]. As can be seen, ω_H and the coefficients for the group factors sum to the value for coefficient omega.

Table 1:
Model Fit of Confirmatory Bifactor Analyses of the GIK 4–6

Model	Description	$\chi2$	*df*	*p*	CFI	RMSEA	SRMR
1	Initial model	44.23	3	.000	.98	.11	.03
2	Final model	1.27	1	.259	1.00	.02	.01

Table 2:
Standardized Loadings of the Final Bifactor Model

Subtest	*g*	Reading	Math & Science
Verbal Ability I	.70	.53	
Verbal Ability II	.49	.25	
Preknowledge in Science	.65	.60	.28
Preknowledge in Mathematics	.61		.41
Nonverbal Ability	.58		

Discussion

The aims of the recent study were to highlight the rarely discussed problem of reliability estimation and interpretation of total scores from multidimensional cognitive measures and to evaluate the cognitive ability section of the GIK 4–6, a newly developed test battery to identify gifted students in the United Arab Emirates, using confirmatory bifactor analyses.

Concerning the factor structure, we found *g*, a reading factor, and a mathematics and science factor to influence the performance in the subtests. This model fit the data extremely well. Coefficient omega of the total score was .89, 95% CI [.87, .91], which means that 89% of the total score variance is due to these factors. It should be noted that the value of .89 most likely underestimates reliability since the test battery includes a single nonverbal ability test for which we could not estimate another group factor, as the nonverbal ability test would have been the only indicator. The reliability of the total score is sufficiently high to use it for selecting students based on the total scores. The proportion of variance of the total score due to the *g* factor, ω_H, was .64, 95% CI [.58, .69]. Regarding interpretation of the total score, this means that the total score is predominantly measuring *g*, although it is not high enough to interpret the total score as essentially unidimensional. The proportion of variance of the total score due to the reading factor was .20, 95% CI [.15, .26] and the corresponding proportion due to the mathematics and science factor was .06, 95% CI [.05, .06]. As can be seen, especially the relatively strong influence of the reading factor, which is uncorrelated with *g*, may not be neglected in interpreting the total score. Therefore, the total score should best be interpreted as a blend of *g*, a rather strong reading factor, and a mathematics and science factor. The reading factor is most likely assessing reading speed since Verbal Ability I and Preknowledge in Science clearly showed the highest loadings – the two subtests for which reading speed seems very important.

Concerning the psychometric evaluation of cognitive ability test batteries in general, we want to highlight the following considerations (for a similar discussion see Brunner & Süß, 2005). The prominent CHC theory explicitly postulates a multidimensional structure of human cognitive abilities, classified by several correlated group factors that summarize subsets of similar abilities. Nonetheless, usually a total score is formed to assess the general intelligence factor *g*, justified by the ubiquitous finding that all kinds of cognitive ability tests (and the group factors) are correlated. However, as there are considerable differences between test batteries, we consider it important for interpretation and reliability estimation to know to what degree the total score of a certain test battery is measuring *g*, and to what degree it is measuring which group factors. In our opinion, confirmatory bifactor analyses should more regularly be used to evaluate test batteries of cognitive ability regarding interpretation and reliability. Additionally, the procedure can also provide valuable information about the appropriateness of subscales and to assess the suitability of a set of items for scaling according to item response theory (Reise et al., 2010).

Limitations

Compared to previous bifactor analyses of cognitive ability test batteries, which analyzed the factor structure based on 45 (Brunner & Süß, 2005), respectively 15 subtests (Gignac & Watkins, 2013), we had a rather low number of indicators, as the cognitive ability section of the GIK 4–6 consists of only five subtests. This limits the possibilities when evaluating the underlying factor structure. For example, according to CHC theory, it seems likely that the subtest Nonverbal Ability, which uses composite figures, is influenced by a visual processing factor (Gv). As the subtest Nonverbal Ability is most likely the only indicator influenced by Gv, it was not possible to test this assumption. Additionally, interpretation of the factor we labeled mathematics and science is rather unclear. Interpreted within CHC theory, it seems to fit best the category labeled by Carroll as "Abilities in the domain of knowledge and achievement" (McGrew, 2009).

Acknowledgments
The study was funded by the Hamdan bin Rashid Al Maktoum Foundation for Distinguished Academic Performance.

References

Beaujean, A. (2015). John Carroll's Views on Intelligence: Bi-Factor vs. Higher-Order Models. *Journal of Intelligence*, *3*(4), 121–136. http://doi.org/10.3390/jintelligence3040121

Brunner, M., & Süß, H. M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*(2), 227–240. http://doi.org/10.1177/0013164404268669

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. http://doi.org/10.1037/0021-9010.78.1.98

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, Dimensionality, and Internal Consistency as Defined by Cronbach: Distinct Albeit Related Concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4–9. http://doi.org/10.1111/emip.12095

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. http://doi.org/10.1111/bjop.12046

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. http://doi.org/10.1080/00273171.2013.804398

Green, S. B., & Yang, Y. (2015). Evaluation of Dimensionality in the Assessment of Internal Consistency Reliability: Coefficient Alpha and Omega Coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14–20. http://doi.org/10.1111/emip.12100

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

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). Berlin Intelligence Structure-Test, Form 4. Göttingen, Germany: Hogrefe.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.

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

Morgan, G., Hodge, K., Wells, K., & Watkins, M. (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. http://doi.org/10.3390/jintelligence3010002

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

Muthén, L. K., & Muthén, B. O. (1998-2010). Mplus user's guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. http://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. http://doi.org/10.1080/00223891.2010.496477

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Social Indicators Research*, *102*(3), 443–461. http://doi.org/10.1007/s11205-010-9682-8

Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - And a little bit more. *Intelligence*, *30*(3), 261–288. http://doi.org/10.1016/S0160-2896(01)00100-3

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625. http://doi.org/10.1007/BF02289858

Wechsler, D. (2008a). Wechsler Adult Intelligence Scale–Fourth Edition. San Antonio, TX: Pearson Assessment.

Ziegler, A. & Stoeger, H. (2016). Gifted Identification Kit 4-6 for the United Arab Emirates. Dubai; UAE: HADAP.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. http://doi.org/10.1007/s11336-003-0974-7