Intelligence 44 (2014) 120–133

Contents lists available at ScienceDirect

Intelligence

Consistent g- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries

Sonja Valerius, Jörn R. Sparfeldt *

Saarland University, Department of Educational Science, Campus A5 4, D-66123 Saarbrücken, Germany

ARTICLE INFO

Article history:
Received 3 March 2013
Received in revised form 16 March 2014
Accepted 8 April 2014
Available online 4 May 2014

Keywords:
Intelligence
g-Factor
Domain-specific ability
Confirmatory factor analysis
Nested-factor modeling

ABSTRACT

Concerning the correlational structure of intelligence, there is a broad consensus regarding hierarchical models with a general factor at the apex (*g*), and less consensus regarding the number, content, and structure of more specific ability-factors hierarchically below *g*. Previous studies revealed very high correlations of test-battery-specific *g*-factors, whereas the consistency of more specific ability-factors has been neglected. In order to investigate this, current data stemming from $N = 562$ high school students who took 26 mental ability tests from independently developed test-batteries were analyzed. Regarding the intelligence-structure, nested-factor models revealed a (relatively) better fit than higher-order models and general-factor-models. The test-battery-specific *g*-factors of the nested-factor models were substantially correlated ($r \geq .91$); the correlations of the test-battery-specific verbal and numerical factors evidenced convergent and discriminant validity (convergent correlations: verbal – $r = .83$; numerical – $r = .46$; figural – $r = .22$). These results provided evidence that some group factors (besides the *g*-factors) of different test-batteries are largely similar.

1. Introduction

There has been a discussion lasting decades about the correlational structure of cognitive abilities, specifically “intelligence”. At present, there seems to be a broad consensus supporting a hierarchical model with a single general factor at the apex (often referred to as [Spearman's] *g*; e.g., Carroll, 1993; Deary, 2012; Jensen, 1998; Spearman, 1904). Nevertheless, there is less consensus regarding the number and content of more specific ability-factors hierarchically below *g*, as well as, the specific structure of these group factors. Focusing on hierarchically structured intelligence conceptions, *nested-factor*-models (sometimes synonymously labeled as *bifactor* models [e.g. Chen, West, & Sousa, 2006; Holzinger & Swineford, 1937; Reise, 2012]) have been introduced as an empirically and theoretically well-interpretable alternative (e.g., Brunner, Nagy, & Wilhelm, 2012; Gignac, 2008; Gustafsson & Balke, 1993; Reise, 2012) to: (a) *higher-order*-models, especially when a general factor *and* domain-specific factors were of interest (Chen et al., 2006), and/or (b) a *g*-factor model without an intermediate hierarchical level of group factors between *g* and specific subtests. Regarding the assessment of intelligence with different test batteries, recent studies revealed very high correlations of the test-battery-specific *g*-factors (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis, & Bouchard, 2008), thereby supporting the consistency of the corresponding *g*-factors. But, the question of the (additional) consistency of more specific ability-factors has been neglected. By combining these aspects, the objective of the present paper is threefold: (1) a comparison of different model specifications within the cognitive abilities' correlational structure, (2) an analysis of the *g*-factors consistency stemming from different test-batteries, and (3) a consistency analysis of different hierarchically intermediate ability factors stemming from different test batteries.

* Corresponding author. Tel.: +49 681 30257490; fax: +49 681 30257488.
E-mail address: j.sparfeldt@mx.uni-saarland.de (S. Valerius).

http://dx.doi.org/10.1016/j.intell.2014.04.003
0160-2896/© 2014 Elsevier Inc. All rights reserved.

1.1. Consistency of g

Although there has been a long history of discussions about the correlational structure of intellectual abilities in scientific psychology, many modern intelligence researchers and intelligence theories agree that "the *g*-based factor hierarchy is the most widely accepted current view of the structure of abilities" (Neisser et al., 1996, p. 81). For example, in his seminal synthesis Carroll (1993) introduced his three-stratum model with a general level (stratum III; similar to *g*), eight broad-level ability factors (stratum II), and even more specific ability factors (stratum I) (see also the Cattell–Horn–Carroll theory of cognitive abilities, CHC; e.g., McGrew, 2009). Focusing on the assessment of intelligence, even a cursory glance at different research projects and intelligence tests reveals a broad range of different tasks to assess different facets of "intelligence", varying in, at least, the number, names, content, and composition of these specific tasks. This relative disagreement regarding the specific number, names, content, and composition of tasks within one test-battery might result in different *g*-factors, or in other words: more or less similar test-battery-specific *g*-factors. Applied psychology relies particularly on the consistency of the measurement of intelligence. Following studies that inspected the nature of the *g*-loadings of particular tests (e.g., Thorndike, 1987; Vernon, 1989), as well as, following studies that compared different factor-analytic methods to extract *g* (e.g., Jensen & Weng, 1994; Ree & Earles, 1991; see also Floyd, Shands, Rafael, Bergeron, & McGrew, 2009; Major, Johnson, & Bouchard, 2011); two recent studies addressed directly the extent to which *g*-factor-scores depend on the specific tasks and abilities assessed (Johnson et al., 2004, 2008).

Johnson et al. (2004) factor-analyzed the data of $N = 436$ adults from three different intelligence test batteries (*Comprehensive Ability Battery*, CAB, Hakstian & Cattell, 1975 [14 tests]; *Hawaii Battery*, HB, DeFries et al., 1974, including Raven's Progressive Matrices, 1941 [17 tests]; *Wechsler Adult Intelligence Scale*, WAIS, Wechsler, 1955 [11 tests]). The fit indices from the three models (one for each of the three batteries), with a second-order *g*-factor at the apex for each, were at least acceptable — in accordance with Hu and Bentler's (1999) cut-off criteria (RMSEA [CAB/HB/WAIS] = .031/.050/.061). The fit of a combined model from three test-battery-specific hierarchical models with three test battery-specific *g*-factors at the apex was acceptable, as well (RMSEA = .069); the correlations of the corresponding *g*-factors ranged from .99 to 1.00. Additionally, this test-battery-specific second-order model showed a better fit than (a) a model with battery-specific *g* factors and without an intermediate hierarchical level (RMSEA = .104; Johnson et al., 2004, p. 104) and (b) a comparable fit for "a model with the same first-order structure … but with only one single *g*-factor" (Johnson et al., 2004, p. 104; RMSEA = .069). Consequently, Deary (2012) concluded that "the individual differences in *g* were identical from the three different batteries" (p. 457). In a replication study, Johnson et al. (2008) reanalyzed a data matrix by de Wolf and Buiten (1963) that was based on $N =$ 500 16-year old Dutch seaman. Thereby, Johnson et al. (2008) used 44 tests from five batteries for their analysis: (1) the *Battery of Royal Dutch Navy* (RDN; 8 subtests), (2) an adaptation of a test battery from the *Twente Instituut voor Bedrijfspsychologie* (TIB; 13 subtests), (3) the *Cattell Culture Fair Test* (CCFT; 4 subtests), (4) the *General Aptitude Test Battery* (GATB; 12 subtests), and (5) the *Groningse Intelligentie Test* (GIT; 7 subtests). Confirmatory factor analyses revealed good to acceptable fit-statistics in separate analyses for each test-battery with a test-battery-specific second-order *g*-factor (RMSEA [RDN/TIB/CCFT/GATB/GIT] = .071/035/.000/.046/.040), as well as, the combined model of all test batteries (RMSEA = .073). Again, the correlations of the test-battery-specific *g*-factors were very high, ranging from .77 for CCFT with GATB to 1.00 for TIB with GATB and GIT, respectively. But, by restricting each test to load only on the corresponding factor of the battery from which it stemmed, some *g*-factor correlations would have exceeded the statistical boundary of 1.00. Thus, Johnson et al. (2008) allowed residuals and first-order factors to correlate across batteries in order to reduce *g*-factor correlations, that rose above 1.00 to 1.00. Summing up the findings: "[t]hese results provide evidence both for the existence of a general intelligence factor and for the consistency and accuracy of its measurement" (Johnson et al., 2008, p. 91). Nevertheless, there were substantial and systematic correlations between subtests and first-order factors not accounted for by the *g*-factors, either.

1.2. Alternative conceptions: higher-order-, nested-factor-, and general-factor models

Whereas there is little doubt about the psychometric existence, generality, stability, and relevance of *g* (e.g., Deary, 2012; Gottfredson, 1997; Jensen, 1998; Jensen & Weng, 1994; Lubinski, 2004), disagreement about the factor structure besides or below *g* is still relatively widespread. In contrast to rather well-known (a) *general-factor models* without a further hierarchy (GF-models, sometimes called one-factor models), especially (b) *higher-order-factor models* (HO-models) and (c) *nested-factor-models* (NF-models) are discussed (e.g., Brunner et al., 2012, see also Gignac, 2005, 2006b; Gustafsson & Balke, 1993). Additionally, these different models correspond with alternative statistical representations to account for the variance–covariance matrix of hierarchically structured ability constructs in the framework of confirmatory factor analysis (CFA). Furthermore, CFA allows a statistical comparison of these different model specifications. The GF-model assumes one general factor (*g*) that summarizes and represents statistically the covariances of the specific subtests. Thereby, individual differences in each specific cognitive task or subtest are influenced by a combination of (a) *g* and (b) a test-specific and *g*-independent additional factor, being a mixture of test-specific reliable variance and error-variance — hence the name "two factor theory". A distinction of reliable test-specific variance and random/error variance within the test-specific variance is not possible. In terms of a more substantive interpretation, these test-specific and *g*-independent variance components are usually assumed to be negligible in the framework of the GF-model.

In contrast, HO-models and NF-models are statistical representations of (more) hierarchically structured intelligence models with additional factors besides *g* and specific tasks or subtests (e.g., Carroll, 1993; McGrew, 2009); currently, both models are assumed to be more adequate statistical representations of the correlational structure of intelligence than the GF-model. In HO-models (left hand side of Fig. 1), similar subtests are assumed to be influenced by one first-order or so-called group factor, whereby the covariances of these group-factors are accounted for by the second order *g*-factor at the apex. Usually the group-factors refer to subtests that consist of items from either a specific content domain (e.g., verbal, numerical) or different operationalizations to assess a medium-specific cognitive operation (e.g., reasoning, memory). In this manner, individual differences in each specific cognitive task are accounted for by a combination of (a) a specific group-factor, and thereby indirectly, (b) *g*, along with (c) residual-terms (consisting of a combination of reliable test-specific variance components and unreliable error variance). These residual-terms are (mostly) specified to be mutually independent (and, then, correlate neither with the group factors nor with *g*). Usually there are residuals of the group-factors not completely accounted for by *g*; and one should keep in mind that each specific task is influenced indirectly (via the corresponding group-factor) and not directly by *g*. In this context Gignac (2008), referring to Yung, Thissen, and McLeod (1999), pointed out that the association between the *g*-factor and the observed variables is mediated completely by the group-factors.

NF-model (right hand side of Fig. 1) refers to a model specification in which the domain-specific ability factors are nested within a more general factor (*g*), whereby *g* and the domain-specific factors, as well as, the different domain-specific factors among themselves are assumed to be mutually uncorrelated. Thereby, individual differences in each specific cognitive task are accounted for by a combination of the following mutually independent factors: (a) *g*, (b) a specific group-factor and (c) residual-terms (consisting of a combination of reliable test-specific variance components and unreliable error variance). Therefore, the *g*-factor as the broadest ability factor is specified as a first-order factor that *directly* influences *all* subtests of a psychometric measure. Therefore one difference to a HO-conceptualization lies in the "broad", but direct (NF) vs. "superordinate", but indirect (HO) conceptualization of *g* (see Gignac, 2008).

1.3. Higher-order- versus nested-factor-models

Within psychometric intelligence research the HO-model seems to be a popular model specification that is often adopted without considering alternatives. However, the NF-model is discussed as an alternative model specification. Before comparing these two models in more detail, one should keep in mind that HO- und NF-models would be mathematically equivalent, if one would add in the HO-model (constrained) direct effects from the general-factor to every observed variable (i.e., test) additionally to the general-factor effects on the domain-specific factors (Yung et al., 1999). Because these effects are usually eliminated in a "standard" HO-model, the HO-model represents a constrained version of the NF-model. Therefore, the NF-model is less restrictive and the HO-model is more parsimonious (and more restrictive because all direct effects of *g* on the observed variables are constrained to zero). Another difference concerns the orthogonality of the first-order factors within the NF-model (among the group-factors, as well as, with the *g*-factor). Because the group factors of the NF-model refer to covariances that are independent of *g*-bound-variances, the group-factors refer to the residuals of *g*. From a statistical point of view the uncorrelated factor-structure in the NF- model enables the (statistically) independent and distinct analysis of domain specific factors *and* a general factor, as well as, their respective correlates (see Chen et al., 2006). When using the NF-model one "can test, less ambiguously, hypotheses pertaining to the existence and nature of factors, beyond the general factor" (Gignac, 2006a, p. 143).

Nevertheless, from a more psychometric perspective, there are questions to be answered regarding the substantive meaning of the factors in NF-models: On the one hand, *g* is based on the covariances of the items (and not a more substantive psychological construct) and on the other, the group factors are based on the covariances of the (*g*-)residuals. In HO-models, however, the substantive meaning of the group factors can be induced out of an inspection of the corresponding tests (e.g., tests that are supposed to tap, for example, "memory"). Nevertheless, *g* in HO-models is based on the covariance of the group factors, as well. Jensen and Weng (1994) argued that the NF-model is not explicitly hierarchical, because *g* does not depend on the variable's loadings on the domain-specific factors. But, the hierarchical dependency in HO-models could also be distracting as mentioned by, for example, Brunner, Nagy and Wilhelm (2012; Schmiedek & Li, 2004; cf. Yung, Thissen and McLeod, 1999): these authors referred to the *proportionality constraints* of

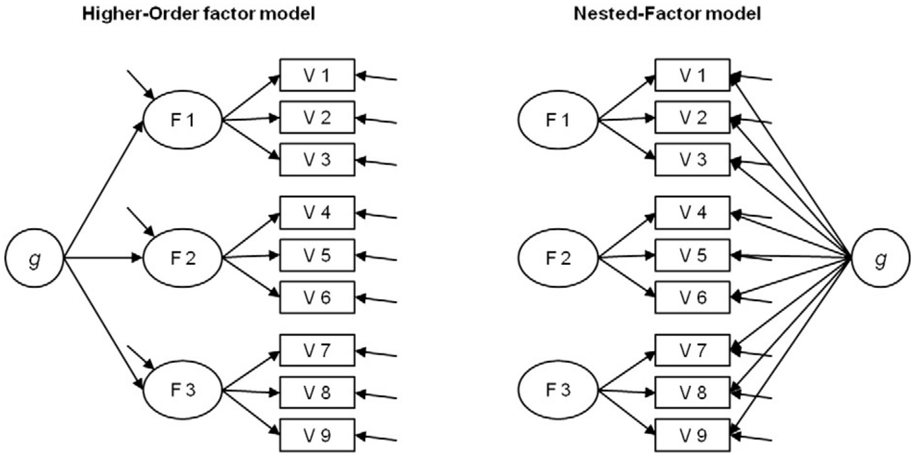

Fig. 1. g = general factor, F1–F3 = domain-specific factors, V1–V9 = manifest variables.

HO-models — meaning that for the indicators of one domain-specific factor the ratios of variance attributable to the respective residual of the domain-specific factor and variance attributable to the general-factor are the same due to the constraint. This statistical constraint in HO-models complicates a straight-forward interpretation for the corresponding correlation coefficients of the group factors and of *g* with other variables. Other authors mentioned substantial advantages of NF-models compared to HO-models (e.g., Chen et al., 2006; cf. Brunner et al., 2013), especially when one is interested in: (a) evaluating the psychometric properties of a test with "a strong common trait [e.g., the *g*-factor], but there is multidimensionality caused by well defined clusters of items from diverse subdomains" (cf. Reise, 2012, p. 692) — as particularly in intelligence tests and/or (b) the unique correlations of the general factor and more domain-specific (group) factors with other variables. This last aspect seems to be especially useful when "making it possible to disentangle differences in the general level of … achievement from differences in specific strengths and weaknesses", specifically in an analysis of the general intelligence level and "particular shapes of [the] performance profile" (Brunner et al., 2013, p. 394).

Some CFA-studies dealt with the construct validity of several intelligence test batteries that provided evidence for the NF-model as a useful and well fitting (alternative) model specification (e.g. Brunner et al., 2012; Gignac, 2005, 2006b; Gustafsson & Balke, 1993). Brunner et al. (2012) compared different model specifications by using data from the Spanish standardization sample of the Wechsler Adult Intelligence Scale (WAIS-III, see Colom, Abad, Garcia, & Juan-Espinosa, 2002). In contrast to the one-factor model with only one *g*-factor loading on all subtests (RMSEA = .13), alternative model specifications showed likewise and good approximations to the empirical data: (1) a first-order structure with four correlated factors (verbal comprehension [VC], perceptual organization [PO], working memory [WM] and perceptual speed [PS]; RMSEA = .07), (2) a HO-model with four first-order factors (VC, PO, WM, PS) together with a second-order *g*-factor (RMSEA = .07), and (3) a NF-model with orthogonal factors (VC, PO, WM, PS, and *g*; RMSEA = .06). The HO-model (with its proportionality constraints) represents a restricted version of the NF-model; χ^2-difference testing (see Yung et al., 1999) evidenced a better fit of the NF-solution. Additionally, the one-factor model revealed a worse fit than the NF-model. Based on these findings, the authors concluded that the NF-model depicts a reasonable representation of the empirical data of the WAIS-III, and that domain-specific ability factors account for a substantial amount of common variance among subtest-scores above the general factor (Brunner et al., 2012, p. 812). Furthermore, Gignac (2005) analyzed the data of the standardization samples of the WAIS-R, again evidencing a (numerical) superiority of the NF-model with three factors (*g*, Verbal-IQ [VIQ], Perceptual-IQ [PIQ]; RMSEA = .06) compared to a GF-model (RMSEA = .12), a HO-model (RMSEA = .09) and an oblique factors model with two factors (VIQ and PIQ; RMSEA = .09). In a replication study with the data of the standardization sample of the WAIS-III (Gignac, 2006b), the NF-solution with three factors (*g*, VIQ, PIQ; RMSEA = .08) revealed again a better fit than the HO-model with a second order *g* and first order VIQ- and PIQ-factors (RMSEA = .11).

Focusing on a comparison of the models, NF-intelligence models often revealed a numerically better fit to empirical data than HO-models (e.g., Brunner et al., 2012; Murray & Johnson, 2013). But this conclusion of a superiority of NF-models based on the fit was criticized by Maydeu-Olivares and Coffman (2006) among others, who reported insufficient power to reject incorrect NF-models in their small simulation study. Focusing on the comparison between HO- and NF-models Chen et al. (2006) evidenced sufficient power for the rejection of incorrect NF-models in their simulation studies. Recently, Murray and Johnson (2013) compared the HO- and the NF-models with data based on 42 tests arranged in two test-batteries with 21 tests each. Whereas the NF-model showed a better fit for both batteries in real data comparisons, an additional simulation study revealed, that the NF-model was favored even when the true model was a higher-order model (p. 419). Therefore, further research is urgently needed regarding the question of power when rejecting incorrect models, as well as, comparing different models, especially in the framework of cognitive ability structure research. Based on these results, Murray and Johnson (2013) concluded that the decision "in the absence of strong substantive or empirical reasons for preferring either model, which is to be preferred may ultimately depend on the purpose of the measurement model" (p. 420).

Indeed, the HO-model has often been selected as an appropriate method. But, if one assumes a strong common trait *and* the multidimensionality of the defined clusters of sub-domains (see Reise, 2012) and accepts a rather "breadth" interpretation of *g* with direct effects on every observed variable, in accordance with Spearman's understanding of *g*, the NF-model represents a statistically suitable alternative to the commonly used HO-model, especially when one is interested in the correlations of *g* and the group-factors with external criteria (see Brunner et al., 2013; Chen et al., 2006; Murray & Johnson, 2013).

1.4. The Berlin model of intelligence structure

A careful inspection of current intelligence concepts that analyze intelligence factors simultaneously at different hierarchical levels revealed the *Berlin Model of Intelligence Structure* (BIS; Jäger, 1982; Süss, Oberauer, Wittmann, Wilhelm, & Schulze, 2002) as especially well-suited. Historically, the BIS was developed following a sample examination of about 2000 different tasks to assess intelligence and different intelligence facets. After eliminating tasks that were doubled or very similar to each other, the bi-factor BIS-structure with three content facets (verbal, numerical, and figural) and four operation facets (reasoning capacity, memory, speed, and creativity), resulting in twelve content-operation-combinations (3 contents × 4 operations; see Fig. 2) in addition to *g* at the apex, was replicated repeatedly (cf. Beauducel & Kersting, 2002). Basically, the model structure can be interpreted as a classification scheme for different intelligence tasks, as well. A verbal analogy task, for example, can be fitted into the BIS cell which results from a cross between the "verbal" content-facet and the "reasoning" operation-facet. Therefore, achievement in a specific task is influenced by a specific *operation*-, a specific *content*- and the *general*-intelligence factor *g* (along with error-terms consisting of a combination of reliable test-specific variance components and unreliable error variance). As mentioned, this

BIS-classification scheme does not just allow a task classification of the BIS-tests developed to assess intelligence following this conception, but also as a more general framework to classify intelligence-tasks in general. Of particular interest is the distinction among different content-facets according to the material used (verbal [V], numerical [N], figural [F]).

NF-modeling with data from the Berlin-intelligence-structure test (Jäger, Süss, & Beauducel, 1997; for an English description see Süss et al., 2002) showed a reasonable fit to empirical data. Specifically, Brunner and Süss (2005, 2007) analyzed NF-models with eight orthogonal factors (three *content*-, and four *operation*-facet-factors along with the *general*-factor *g*) in a CFA (RMSEA = .04) in which the tests were restricted to load only on their respective content- and operation-facet factors, as well as, on the general factor. Additionally, two separate NF-models with (a) only the content-facet-factors (RMSEA = .03) besides a general factor or (b) only the operation-facet-factors (RMSEA = .04) besides a general-factor revealed in both cases a good model fit, as well.

1.5. Consistency of group factors

Whereas there is empirical evidence for the consistency of *g* (as outlined above), the (additional) analysis of the specific ability factors consistency besides *g* has been neglected. Focusing on the consistency of *g*, Johnson et al. (2008) did not inspect the covariances of these specific ability factors systematically. Nevertheless, these authors concluded that "[t]here are substantive correlations among … specific abilities from battery to battery, and from first-order factor to first-order factor, and different tests measure them with reliability comparable to that associated with the general factor" (p. 91). At least to our knowledge, an additional and more systematic investigation of these "besides-*g*-covariances" of the corresponding "besides-*g*-factors" of different test-batteries is still to come. Theoretically, the outlined characteristics of NF-models seem to be especially well-suited to answer such a research question because one could inspect the consistency of *g* and the consistency of more specific group factors simultaneously. Specifically, NF models allow the estimation of group factors and their cross-battery correlations free of *g*-variance.

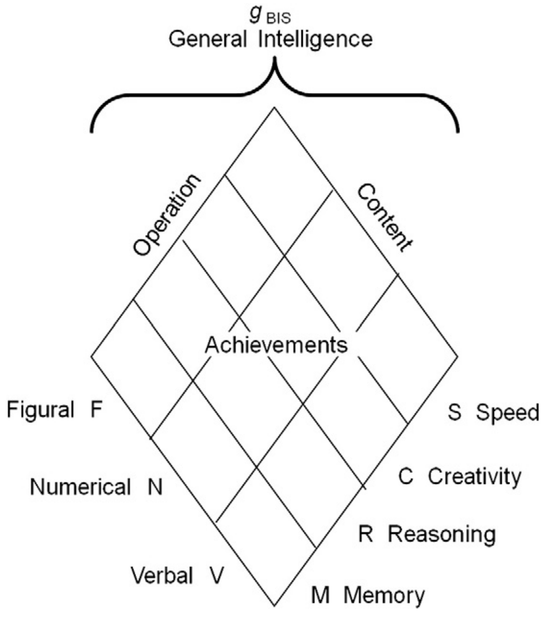

Fig. 2. The BIS model (Jäger, 1982).

1.6. The present study

Taking these aspects into account, the aims were threefold of the present study investigating different structure facets of intelligence tasks stemming from three different and independently developed test-batteries: (1) following prior analyses (e.g., Brunner et al., 2012; Gignac, 2005, 2006a,b; Gustafsson & Balke, 1993), different hierarchical models of cognitive abilities (GF-model, HO-model, NF-model) were compared in the framework of confirmatory factor analyses. The GF-model (as a kind of "baseline-model") represents a rather broad conceptualization with a strong common influence of the *g*-factor. We expected an increasing model fit with specifying additional group-factors. Based on prior findings, we expected specifically (a) at least acceptable absolute model-fit indices of all three models and (b) a relatively better fit of NF-models (and partially HO-models) than GF-models. This increasing model fit would correspond with a substantially meaningful and, therefore, more adequate representation of systematic covariances besides the general factor. (2) Conceptually replicating the studies by Johnson et al. (2004, 2008), recent data were collected and analyzed in order to investigate the consistency of *g*-factors stemming from three different test-batteries. Based on prior results, we expected (very) high positive correlations of these test-battery-specific *g*-factors. (3) Over and above prior findings, the consistency of group factors stemming from three different test-batteries was to be investigated in a NF-model structure. Importantly, the structure of the general *and* of the domain-specific factors across batteries could be taken into account. We expected (besides [very] high positive correlations of these test-battery-specific *g*-factors) at least substantial positive correlations of these corresponding test-battery-specific intermediate factors.

2. Method

2.1. Participants

Participants were $N = 562$ German academic-tracked high school students (*Gymnasium*, grade 9) from 23 classes out of six schools ($n = 279$ females, $n = 258$ males, $n = 25$ without gender specifications; mean age = 15.6 years, $SD = 0.45$, Min = 14.2, Max = 17.4). The participation rate was 91%; the parents of 9% of the high school students did not allow their child to take part in this investigation.

2.2. Instruments and procedure

A description of the 26 tests, their allocation to a specific content facet (verbal, numerical, figural), the time limits, and the number of items is presented in Table 1. These 26 tests stemmed from well-known and widely used German intelligence test-batteries and formed three (new) test-batteries.

Table 1
Tests included in the batteries, supplemented by a description of the tests, the corresponding content-facet, the allocated allowed time, the number of items, the means (M), the standard deviations (*SD*) and the intraclass-correlations (ICC).

	Test (abbr.)	Description	Mat. factor	Time allowed	No. of items	M	*SD*	ICC
Berlin Intelligence Structure Test, Form 4 (BIS-4), short version								
1.	City map (OG)	Recall of buildings in a city map	F	1:30 + 1:40	27	15.78	4.24	.06
2.	Number sequences (ZN)	Completion of numbers in a series	N	3:40	9	5.25	2.17	.10
3.	Relevant traits (EF)	Generation of traits for a special occupational group	V	2:30	f.r.	10.14	3.71	.07
4.	Figural analogies (AN)	Identification of analogous figure to follow a sequence of figures	F	1:45	8	3.97	1.63	.16
5.	x greater (XG)	Crossing out numbers x greater than the prior one	N	1:00	44	22.03	8.11	.06
6.	Verbal analogies (WA)	Identification of analogous word pairs	V	1:30	8	3.30	1.67	.07
7.	Layout (LO)	Shaping of graphical labels	F	3:00	f.r.	4.27	1.48	.09
8.	Paired associates (ZP)	Recall of numbered pairs	N	2:00 + 2:00	12	7.07	2.54	.08
9.	Fact-opinion (TM)	Conclusion of fact or opinion of verbal statements	V	1:00	16	9.30	3.17	.07
10.	Crossing out letters (BD)	Crossing out letters in an arrangement of several ones	F	0:50	130	55.36	10.88	.09
11.	Estimation (SC)	Estimation of complex arithmetic	N	2:45	7	4.15	1.70	.07
12.	Story (ST)	Recall of text information	V	1:00 + 2:00	22	10.60	3.51	.05
13.	Divergent computation (DR)	Generation of arithmetic with given elements	N	1:50	f.r.	11.40	4.22	.05
14.	Charkow (CH)	Completion and generation of figures in a series	F	3:00	6	2.49	1.44	.06
15.	Part-whole (TG)	Crossing out words in a series of words	V	0:40	22	14.15	3.25	.05
Fluid battery with Cattell's Culture Fair Test, Scale 2, short version and Raven's Standard Progressive Matrices								
16.	Series (RF)	Identification of the next element in a series	F	4:00	15	12.79	1.53	.08
17.	Classification (KL)	Identification of the element in each group that does not belong	F	4:00	15	10.59	2.17	.03
18.	Matrices (MZ)	Identification of the analogous element of the matrix	F	3:00	15	11.86	1.93	.06
19.	Conditions (topology) (TO)	Identification of the topologically equivalent element	F	3:00	11	7.05	2.01	.04
20.	SPM set A–E	Identification of the analogous element of the matrix	F	45:00	60	50.98	4.29	.09
German Cognitive Ability Test for 4th–12th grades, short version								
21.	Vocabulary (WS)	Identification of words with similar or same meaning	V	7:00	25	17.53	3.01	.06
22.	Verbal analogies (WL)	Identification of analogous word pairs	V	7:00	20	12.07	2.77	.06
23.	Quantity comparison (MV)	Comparison of greater/smaller relation of numerical elements	N	10:00	25	16.98	3.51	.11
24.	Number sequences (ZR)	Completion of numbers in a series	N	9:00	20	17.51	2.51	.04
25.	Figure classification (FK)	Identification of matched figures	F	9:00	25	21.79	2.66	.07
26.	Figure analogies (FA)	Identification of analogous figure pairs	F	8:00	25	20.62	3.20	.03

Note: f.r. = free response. The participants should generate free responses so that an indication of the number of items is not possible.

Practically, these 26 tests were compiled to form three test-booklets (booklets B 1, B 2, B 3) and were administered booklet-wise in whole school classes, whereby administering a booklet lasted between 50 and 90 minutes. The three booklets were administered in a randomized order for the school classes in the three testing sessions on three different days with an intermediate time period of two to three days. All data were collected during regular lessons by trained experimenters. For few students who could not attend to the regular group testing sessions (because of, e.g., sickness), separate testing was scheduled in smaller groups. Because unfortunately not every high-school student participated in each of the three testing sessions due to illness or for other reasons not specifically related to the study, we were able to collect data at all three measuring points for 87.5%. 11.7% took part at two sessions and 0.7% just at one testing-session. Data collection took place from June to July in 2010.

Booklet 1 (B 1) was comprised of the German adaptation of Cattell's Culture Fair Test (CFT; Weiss, 2006) consisting of four figural tests and the short form of the Berlin Intelligence Structure Test (most recent form: BIS-4; Jäger et al., 1997) consisting of 15 heterogeneous tests, which are briefly described in Table 1. The four tests from the CFT are quite similar in regard to content and cognitive operation; exclusively figural material is presented and the items are supposed to assess (primarily) reasoning. In the (German) test manual, retest reliability coefficients for a period of two to five months were reported for the subtests ranging from r_{tt} = .48 for classification to r_{tt} = .65 for matrices and for the sum value of r_{tt} =.92 (based on the four subtests). The tests of the BIS-4 (short form) assessed the content facets verbal, numerical, and figural with five tests each. Additionally, the sum of all 15 tests is an indicator of g_{BIS}. Unfortunately, stability coefficients of the short version and/or specific tests were not reported in the test manual, a one-year stability coefficient of the general factor of the long version reached r_{tt} = .88 (Süss et al., 1991). Booklet 2 (B 2) consisted of Raven's Standard Progressive Matrices (SPM; Raven, 1941) with figural reasoning tasks. There is considerable evidence for the SPM to be highly *g*-loaded (e.g. Jensen, 1998). In the (German) test manual, the three month retest reliability coefficients reached r_{tt} = .90 for the whole sample (Heller, Kratzmeier, & Lengfelder, 1998). Booklet 3 (B 3) consisted of the short form of the German adaptation of the Cognitive Ability Test (CogAT; Heller & Perleth, 2000) with six tests. Combined retest- and parallel-test reliabilities (three weeks) of the tests (of the long test version) ranged from r_{tt} = .76 for verbal analogies to r_{tt} = .92 for figure analogies. The CogAT assesses – besides g_{CogAT} – verbal, numerical, and figural reasoning (Heller & Perleth, 2000).

2.3. Data analyses

Test scores for each test were computed, serving as basis for the further analyses. The 26 tests formed three test-batteries: the tests stemming from the BIS-test formed the BIS-battery and the tests stemming from the CogAT formed the CogAT-battery.

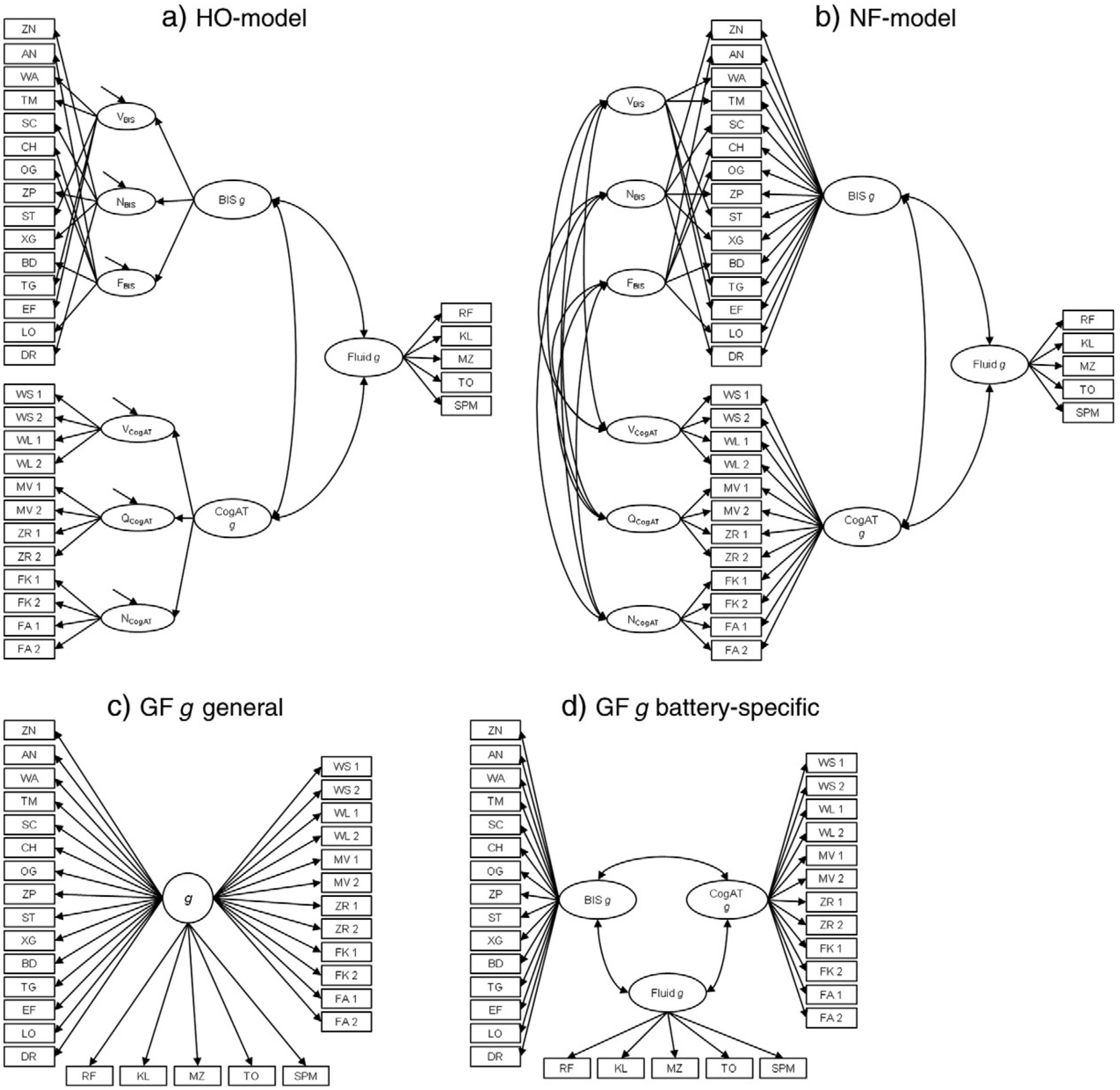

Fig. 3. Confirmatory factor analyses for alternative models estimating general and specific factors in three test batteries. *g* = general ability, BIS = Berlin Intelligence Structure Test, Form 4, CogAT = German Cognitive Ability Test for 4th to 12th grades, Fluid = Culture Fair Test, Scale 2 with Raven's Standard Progressive Matrices. V_{BIS} = verbal content factor sensu Jäger's (1982) BIS-model, N_{BIS} = numerical content factor, F_{BIS} = figural content factor. V_{CogAT} = Verbal Reasoning, Q_{CogAT} = Quantitative (numerical) Reasoning, N_{CogAT} = Nonverbal (figural) Reasoning. For descriptions of the manifest variables (subtests) see Table 1. To ensure the clarity of presentation, disturbance terms (all uncorrelated) of the manifest variables are not displayed.

These two batteries included a rather broad range of different cognitive tasks, regarding different content facets (in the BIS-terminology). The remaining subtests stemming from the CFT were combined with the SPM to form an additional third test-battery with a focus on figural reasoning tasks (fluid battery), thereby emphasizing this facet (and its potential relevance for the different *g*-factors).

Concerning the first research question, confirmatory factor analyses (CFAs) were conducted using the software program Mplus (Muthén & Muthén, 1998–2007) in order to evaluate and compare different hierarchical model specifications of cognitive abilities (see Fig. 3): (a) HO-model with test-battery-specific higher order factors and test-battery-specific *g*-factors (HO-model), (b) NF-model with test-battery-specific *g*-factors and nested factors (NF-model), (c) GF-model with one (general) *g*-factor (*g* general) for all individual tests, and (d) GF-model with test-battery-specific *g*-factors (*g* battery-specific). In HO-models, each test-battery-specific higher order factor was indicated by the corresponding content facet-specific-tests as specified in the test manuals (see also Table 1). Similarly in NF-models, each test-battery-specific (and, thereby, content-facet-specific) nested factor was indicated by the corresponding content-facet-specific tests as specified in the test manuals (see Table 1). Regarding the BIS-tests, we specified a model with

Table 2
Fit indices for the alternative models: higher-order model (HO), nested-factor model (NF) and both general factor models (GF: *g* general and *g* battery-specific see Fig. 3).

Model	χ^2	*df*	*p*	SRMR	RMSEA	BIC	AIC
HO	1203.807	461	<.001	.069	.054	76,057.075	75,628.257
NF	899.468	431	<.001	.055	.044	75,898.279	75,339.515
GF (*g* general)	1530.051	464	<.001	.071	.064	76,423.261	76,007.437
GF (*g* battery-specific)	1471.523	461	<.001	.069	.062	76,370.952	75,942.134

Note. SRMR = Standard Root Mean Square of Residuals, RMSEA = Root Mean Square Error of Approximation, BIC = Bayesian Information Criterion, AIC = Akaike Information Criterion.

three content-facet specific factors (V_{BIS}, N_{BIS}, F_{BIS}) and a general factor (BIS *g*) in contrast to a model with four operational components or a combined model with both content and operational components. Similar factors were specified for the CogAT (V_{CogAT}, Q_{CogAT}, N_{CogAT}; CogAT *g*). We used nonstandardized test scores as indicators for the latent factors in HO- and NF-models. In order to avoid local under-parameterizations for the CogAT with two indicators for each latent factor (Vocabulary [WS] and Verbal analogies [WL] for "Verbal", Quantity comparison [MV] and Number sequences [ZR] for "Quantitative", and Figure analogies [FA] and Figure classification [FK] for "Figural") and corresponding identifying constraints, we first ranked the items of each test in the order of the item difficulty; secondly, we split the items making use of the *odd–even-method*. We therefore picked the first, third, fifth (and so on) item to create the "odd"-parcel and picked the second, fourth, sixth etc. item to create the "even"-parcel, both consisting of ten to twelve items each. Because the tests were speeded and presented with increasing difficulty the first items were solved more often than the latter test items. By making use of the *odd–even*-method, we created (as much as possible) equal weighted item parcels. To take into account that those two parcels stemmed from the same subtest, we fixed their non-standardized loadings on the corresponding latent factor(s) to be equal.[1] Each factor variance in each model was set to 1.

To control for potential effects due to the school class affiliation and the clustering of the data (students in classes), analyses were run using the method "complex" available in Mplus after inspecting the intraclass correlations. There were few missing values in each test (Mean = 2.5%; Median = 2.5%; Min = 2.5%; Max = 6.2%), so the analyses relied on the Full Information Maximum Likelihood (FIML) estimation provided in Mplus (Muthén & Muthén, 1998–2007). The parameters were estimated by using the robust maximum-likelihood algorithm (MLR).

While evaluating the models, absolute and relative model fits were inspected. In addition to the χ^2-test, which becomes more sensitive to small model deviations with increasing sample size, we used several descriptive fit-statistics for the model evaluation. Following Hu and Bentler (1998), we used the Standardized Root Mean Square of Residuals (SRMR; Bentler, 1995) together with the Root Mean Square Error of Approximation (RMSEA; Steiger, 1990) to evaluate the goodness of fit. According to Hu and Bentler's (1999) cutoff criteria, an acceptable model fit is indicated by SRMR close to .08 and RMSEA close to .06. Admittedly, these values are more or less accepted guidelines. Moreover, we used the magnitudes of the standardized factor loadings with values greater or equal to .30 being substantial to evaluate the interpretability of the models (see Carroll, 1993; McDonald, 1999). Additionally, the different models were also compared by making use of Schwarz' (1978) Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), preferring models with lower values, and a comparison of the fit value change. Although rather an ad hoc guideline to evaluate the fit difference, and therefore to be used with caution, Chen (2007) suggested that the more complex model should be chosen when the RMSEA difference exceeds .015.

Concerning the second research question dealing with a consistency analysis of the *g*-factors stemming from different test-batteries, we inspected the latent *g*-factor correlations of the different models from the first research question. This was done when the fit of the corresponding models was at least acceptable. Regarding the third research question dealing with a consistency analysis of the hierarchically intermediate factors stemming from different test-batteries, the content-facet factor correlations of NF-models were inspected. Thereby, we relied on criteria regarding a convergent and discriminant validation introduced within the multitrait–multimethod framework (Campbell & Fiske, 1959). Because of a lack of generally accepted scientific criteria regarding the absolute magnitude of the corresponding correlation coefficients, we oriented ourselves on the following criteria: In order to evidence convergent validity, the cross-battery correlations of the corresponding content-specific factors had to differ from zero and approach a "large" effect size (i.e., $r = .50$; see Cohen, 1988). In order to prove discriminant validity, these convergent correlation coefficients had to exceed numerically the remaining (discriminant) cross-battery correlation coefficients of the non-corresponding content-specific factors.

Reliabilities of the content-facet factors in the NF-model were computed making use of the model-based reliability estimate called *omega hierarchical* (ω_h) (cf. Brunner et al., 2012). Omega hierarchical was computed as the ratio of variance attributable to the content specific factor (i.e., the squared sum of the factor loadings of the manifest variables on the associated content specific factor) to the total variance of this scale score (i.e., the sum of the manifest variables' factor loadings squared over the associated content specific factor and the *g*-factor plus the sum of the residual variances of these manifest variables). By means of the variance

[1] As an alternative strategy, we also ran analyses using raw items and setting their loadings to equal. Because those analyses resulted in, for example, correlation coefficients that exceeded the boundary of $r = 1$ (indicating a severe misspecification), that alternative strategy did not seem to be useful nor fruitful.

Table 3
Standardized factor loadings in the combined nested factor-model (NF-model; see Table 2).

Test	BIS-4				CogAT				Fluid
	λ_{VBIS}	λ_{NBIS}	λ_{FBIS}	$\lambda_{BIS\ g}$	λ_{VCogAT}	λ_{QCogAT}	λ_{NCogAT}	$\lambda_{CogAT\ g}$	$\lambda_{Fluid\ g}$
TM	.45*			.25*					
EF	.25*			.08					
WA	.28*			.36*					
TG	.45*			.14					
ST	.37*			.18*					
ZN		.34*		.45*					
XG		.51*		.27*					
ZP		.34*		.14*					
SC		.26*		.42*					
DR		.22*		.17*					
OG			.31*	.35*					
AN			.06	.55*					
LO			.32*	.09					
BD			.39*	.14*					
CH			−.14*	.59*					
WS 1					.51^{a*}			.31*	
WS 2					.55^{a*}			.19*	
WA 1					.43^{b*}			.42*	
WA 2					.42^{b*}			.39*	
MV 1						.06^{c}		.48*	
MV 2						.06^{c}		.56*	
ZR 1						.66^{d*}		.47*	
ZR 2						.63^{d*}		.40*	
FK 1							.49^{e*}	.51*	
FK 2							.53^{e*}	.53*	
FA 1							.23^{f}	.67*	
FA 2							.26^{f}	.58*	
RF									.52*
KL									.44*
MZ									.41*
TO									.42*
SPM									.62*
ω_h	.39	.29	.12	.67	.46	.25	.24	.85	.60

Note. For indications with superscript letters (a–f) the unstandardized factor loadings, each with the same letter, were fixed to be equal. The model-based reliability estimates omega hierarchical (ω_h) for the content specific factors are represented in the bottom row.
* $p < .05$.

attributable (solely) to the general factor in the denominator it becomes clear, that ω_h, and thus the reliability of the content specific factors, decreases as the influence of the general-factor increases.

3. Results

The mean values, standard deviations and intraclass correlations of the tests are presented in Table 1. The intraclass correlations ranged from ICC = 0.03 for classification (CFT test KL) to ICC = 0.16 for figural analogies (BIS-4 test AN).

3.1. Absolute and comparative model evaluation

The specified HO-model with correlated *g*-factors for the three test batteries revealed at least acceptable fit statistics (see Table 2 and Fig. 3). All first-order standardized factor loadings with their corresponding factor were statistically significant ($p < .05$), and many of the standardized factor loadings were substantial with values greater than or equal to .30 (exceptions: LO = .12, BD = .18, EF = .28, DR = .28 and ZP = .29). The second-order standardized factor loadings ranged from .58 for V_{BIS} on BIS *g* to .98 for F_{BIS} on BIS *g* for the second-order structure.

The NF-model with three correlated *g*-factors and six domain-specific factors (V_{BIS}, N_{BIS}, F_{BIS}, V_{CogAT}, Q_{CogAT}, N_{CogAT}) correlated across, but not within batteries revealed good fit indices (Table 2). Most of the standardized factor loadings were statistically significant (Table 3). For 14 out of the 32 manifest variables (44%) substantial factor loadings on both the corresponding battery-specific *g*-factor and the corresponding domain-specific factor occurred. A total of 16 tests (50%) revealed substantial loadings either on their corresponding *g*-factor or on the corresponding specific factor. Two variables (6%) did not load substantially on either the corresponding *g*-factor or on the specific corresponding content-facet factor. In sum, 94% of the tests had at least one substantial loading. Overall, both the fit indices and the inspection of the loadings indicated that the proposed NF-model represents an, at least, acceptable approximation to the empirical data.

Two GF-models, a GF-model with test-battery-specific *g*-factors (*g* battery specific) and a model with only one *g*-factor (*g* general), were additionally specified as a basis for further model-comparisons with a widely used “baseline model”. In contrast to the model-fit of the HO- and NF-model, the model fit for both *g*-factor models was acceptable, but numerically poorer. The standardized factor loadings of the tests were statistically significant in both models ($p < .05$).

Table 4
Correlations of the *g*-factors in the different models (see Table 2): higher order models (HO), nested factor models (NF), and test-battery-specific *g*-factor models (GF) across the three test batteries.

Model	CogAT *g*			Fluid *g*		
	HO	NF	GF	HO	NF	GF
Bis *g*	.92*	.91*	.80*	.99*	1.00*	.91*
CogAT *g*				.95*	.92*	.88*

* $p < .05$.

They ranged from LO = .15 on BIS *g* to FA 1 = .69 on CogAT *g* in the test battery-specific GF-model and from Lo = .12 to FA 1 = .67 for the model with one *g*-factor. Therefore, the fit of both of these models was acceptable, but did not account for the empirical data in a totally sufficient manner.

In sum, all four models revealed acceptable to good model-fits. The NF-model revealed the lowest AIC- and BIC-values, followed by the HO-model; both GF-models revealed substantially higher BIC-values. Regarding ΔRMSEA, the NF-model showed a lower value than the HO-model (ΔRMSEA = .010) and both GF-models (ΔRMSEA = .018 and .020), indicating a better fit of the NF-model. Comparing the HO-model with the other models the RMSEA did not differ by more than .015 (although it was numerically higher for the HO-model than for the NF-model and lower for the HO-model than for both GF-models). The SRMR, the AIC and the BIC-values and additionally ΔRMSEA – as a rule of thumb and, therefore, to handle with caution – indicated a (numerically) better fit of the NF-model compared to the HO-model and to both GF-models, and a better fit of the HO-model than both GF-models.

3.2. Correlations of the g-factors

An inspection of the correlations of the latent *g*-factors of the three different test-batteries within the NF-framework revealed very high coefficients (r = .91–1.00; Table 4). Model-based reliability estimates ω_h of the general factors reached .67, .85 and .60 for BIS, CogAT and Fluid, respectively. The *g*-factor correlations were of comparable magnitude within a HO-framework (r = .92–.99). However, even for the less convincing GF-model with test battery-specific *g*-factors (regarding the fit values as well as the model comparisons), the correlations of these *g*-factors stemming from three different test batteries reached substantial values between r = .80 and r = .91.

3.3. Correlations of the content facet-factors

The consistency of hierarchically intermediate intelligence factors (below *g*) stemming from different test-batteries was investigated within the NF-framework. As mentioned, the selected tests of the BIS as well as the CogAT allowed the specification of nested and *g*-independent content-facet factors (verbal, numerical, figural). An inspection of the correlational pattern of these factors revealed a substantial convergent correlation of the two test-battery-specific nested verbal factors (r = .83; Table 5) that was numerically higher than the corresponding (absolute values of the) discriminant correlation coefficients, evidencing convergent-discriminant-validity. For the nested numerical factors, the convergent correlation was statically significant, as well (r = .46), and numerically higher than the absolute values of the corresponding divergent correlation coefficients. Nevertheless, this convergent correlation coefficient was numerically lower than the convergent coefficient for the verbal factors. Regarding the figural nested factors, the convergent correlation coefficient was even lower (r = .22). Moreover, some of the corresponding divergent coefficients were numerically higher in absolute values indicating a lack of convergent–divergent validity of the figural group factors. Model-based reliability estimates ω_h of the content-specific factors ranged from .12 (for *Figural* of the BIS-battery) to .46 (for *Verbal* of the CogAT-battery) (see bottom row of Table 3). These relative low values of the content-specific factors were mainly attributable to the ratio of variance of the general factor in all scales of the BIS- and CogAT-battery. In sum, the NF-model including simultaneously orthogonal general factors and (nested) content-facet-specific factors for each of the independently developed test batteries revealed a distinct correlation pattern for the verbal and numerical group factors over and above the correlations of the general factors.

4. Discussion

The main results of the present study are threefold, whereby the analyses were based on current data from a large sample of high-school students taking a total of 26 heterogeneous intelligence tests stemming from independently developed test-batteries: (1) a comparison of different hierarchically structured intelligence models in a confirmatory approach revealed (a) at least acceptable absolute model fits of the analyzed models (NF-model, HO-model, GF-models with test-battery-specific *g*-factors as well as one general *g*-factor), and (b) a numerically better approximation to the data in the NF-model. (2) The correlations of the three test-battery-specific *g*-factors were very high, indicating that the interindividual differences of the corresponding three test-battery-specific *g*-factors were (almost) interchangeable. This applies to all three methodological approaches (NF-model, HO-model, GF-model with test-battery-specific *g*-factors). (3) Going beyond these and prior findings, the correlations of the content-facet- and besides-*g*-factors evidenced a pattern of convergent–divergent validity of the verbal and numerical group factors in a NF-model framework.

Regarding the first research question dealing with the more specific structure of intelligence, different theoretically derived hierarchical intelligence-models were compared. Corresponding to the majority of theoretical and empirical assumptions (first mentioned by Spearman, 1904, and continued by numerous researchers, e.g., Carroll, 1993; Deary, 2012; Jensen, 1998; McGrew, 2009) all of our CFA-models converged in the notion

Table 5
Correlations of the content-facet-specific factors in the nested-factor models (see Table 2) across the three test batteries.

Factor	V_{CogAT} verbal	Q_{CogAT} numerical	N_{CogAT} figural
V_{BIS}	.83*	.01	.03
N_{BIS}	−.41*	.46*	−.25
F_{BIS}	−.47*	.00	.22

* $p < .05$.

of a general intelligence factor at the apex. Replicating prior results (e.g., Brunner et al., 2012) with data based on a substantially different sample of high school students and conceptually different intelligence tests, our analyses revealed at least acceptable absolute fit statistics for all four model-specifications. Thereby, the widespread assumption of a general factor underlying cognitive ability tasks was supported again. Furthermore, the model fit successively improved from a more general (GF-model) to a more differentiated structure within the HO- and NF-models, that differentiated the latent group-factor structure besides *g* with regard to verbal, numerical and figural content facet factors within the test batteries. The successively better fit of the NF-model than the HO-model and the GF-models is in accordance with the analyses using the Spanish standardization data of the Wechsler Adult Intelligence Scale (WAIS) by Brunner et al. (2012) and the analysis with the standardization samples of the WAIS-R and WAIS-III (Gignac 2005, 2006b). Although the specific NF-model-structure differed somewhat due to different theoretical conceptions between the WAIS with rather operation-facet specific nested factors and in contrast, the content-facet nested factors of our analyses. It would be interesting to (re-)analyze further test batteries in NF-models with a realization of test-classifications according to the theoretically convincing content-facet-specific BIS-facet-structure (already a dimension of Guilford's [1985] Structure-of-Intellect Model) or even the bi-faceted classification schema of diverse intelligence test tasks.

One anonymous reviewer justifiably asked how can one interpret the orthogonality of the factors in the NF-model in terms of human cognitive function, specifically when individual differences are a sum of two (or more) uncorrelated processes. We agree that CFA-models should be consistent with theoretical assumptions. Both NF-models and HO-models are statistical models that represent the empirical data, and for the most part, in terms of model fit at least acceptable. In most cases, a psychological interpretation of the group factors in HO-models seems to be quite straight forward: for example, different tasks and tests are supposed to tap “working memory”, a dimension on which individuals differ. One conclusion is that these inter-individual differences in these specific tasks and tests could causally result from the variation in working memory. Therefore, calling the corresponding group factor a “working memory factor” seems straight forward (thank you to the reviewer for this specification). Nevertheless, the covariances of such group factors in a HO-model constitute what many researchers call *g*; and, therefore, there are purely statistical aspects constituting *g* (taking a more psychological point of view, this might be one reason for the psychological lack of clarity of *g* in HO-models). Similarly, the representation of orthogonal factors in NF-models is “just” a statistical representation of empirical data. One should keep in mind that there are still unresolved issues regarding power and the interpretation of the corresponding fit values change when comparing the models statistically. Taken together, some arguments remain for and against both statistical models. As already mentioned in the Introduction, Murray and Johnson (2013) concluded that the corresponding decision for a particular model specification “may ultimately depend on the purpose of the measurement model” and added: “If ‘pure’ measures of specific abilities are required then bi-factor model factor scores should be preferred to those from a higher-order model” (p. 420). Therefore, from a purely statistical viewpoint and for the purpose of our study, the orthogonality of *g*- and content-specific factors within the NF-specification constitutes an advantage, specifically the distinct analysis of correlations with external criteria (in the present study: content-specific factors from other test-batteries) independently of the general factor and without the inevitable limitations related to proportionality constraints (cf. Chen et al., 2006; Schmiedek & Li, 2004). It would be an interesting challenge for future research to further clarify the relations between more psychological intelligence models and their corresponding statistical representations (cf. Brunner et al., 2013).

Regarding the second research question dealing with the consistency of the *g*-factors stemming from different test batteries, the test-battery-specific *g*-factors correlated very high, as expected. The *g*-factor correlations in the GF-model with one *g*-factor for each battery were substantially lower ($.80 \leq r \leq .91$) in contrast to the HO- ($.92 \leq r \leq .99$) and the NF-model ($.91 \leq r \leq 1.00$). One rather substance-based interpretation of this pattern suggests that the group factors of NF- and HO-models accounted for systematic covariances besides *g*, which somehow biased the *g*-factors of the GF-model. Our results are in line with previous research: (a) the (almost) test battery-independence and consistency of *g* (Johnson et al., 2004, 2008), and (b) a large extent of invariance in the specific factor-analytic method (e.g., Jensen & Weng, 1994). As a test of robustness and to strengthen the latter point we additionally conducted CFA's with all three test-batteries in a combined model and systematically varied the specification types (i.e., NF, HO, GF) for each battery (BIS-HO with CogAT-NF and vice versa; or BIS-GF with CogAT HO and vice versa; and all together with Fluid-GF): all six model-variations showed good to acceptable fit statistics (RMSEA < .06) and the *g*-factor correlations ($r > .84$) were of similar magnitude as in the combined models using uniform methods for the BIS- and CogAT-batteries. It should be kept in mind that the 26 different tests were quite heterogeneous and that the original test-batteries were based on substantially different theoretical intelligence conceptions. These very high *g*-correlations with different tests in an independent sample (German high school students) provided additional evidence for the consistency of *g*.

Of particular importance is the substantial correlation of the battery-specific-*g*-factor of the Fluid-test-battery with both the CogAT- and the BIS-battery-specific *g*-factors. The subtests of the Fluid-battery consisted exclusively of figural reasoning tasks like matrices and figural analogies and were, thereby, capturing a cognitive ability which is rather narrow in scope; the *g*-correlations to the BIS- and CogAT-battery, which are broader in scope, were also very high, ranging from .88 for Fluid-*g* with CogAT-*g* in the GF-model to 1.00 for Fluid-*g* with BIS-*g* in the NF-model. Johnson et al. (2008) reported their lowest *g*-factor-correlation of .77 for the CCFT- with the GATB-battery interpreting this result pattern by suggesting that the CCFT might be narrower in scope, compared to the rather broad GATB-battery. This might have been caused by the fact, that the Fluid-battery in our analyses consisted of (in addition to the four CFT-tasks) the reliable full set of Raven's Standard Progressive Matrices (Raven, 1941) with relatively strong *g*-factor-loadings of .62. Raven's Matrices are known to be (very) highly *g*-loaded (cf., Jensen, 1998). In accordance with prior research and due to our findings, we suggest that the amount of *g*-factor correlations across batteries, although always very high and approaching unity, seems to be influenced by several aspects: (a) the number and/or (b) the differences of the tasks of the batteries, and/or (c) the degree to which the subtests are "*g*-reliable", meaning that high *g*-loaded subtests (e.g. [Raven's] Matrices) strengthen the *g*-factor correlations. Another explanation for the very high correlations between the fluid- and the other batteries could be that we operationalized *g* quite "figural", as indicated by the high loadings of the figural subtests on the *g*-factors in the HO-models of the BIS- (.98), as well as, the CogAT-battery (.84). This point will be particularly important within the framework of the next research question.

The third research question dealt with the inspection of the group-factor correlations of different test-batteries beyond the correlations of test-battery-specific *g*-factors. As expected, the substantial positive correlations of these test-battery-specific and content-facet-specific factors proved (at least for the verbal and numerical factors) convergent validity. Additionally, the pattern of the correlations of those test-battery-specific group factors proved divergent validity. The highest convergent correlation of the test-battery-specific factors occurred for the verbal-group factors ($r = .83$), followed by numerical ($r = .46$) and figural factors ($r = .22$) in the NF-model. As mentioned above, our *g*-factors had a figural focus which could explain (a) the high loadings of the figural factors on the *g*-factors in the higher order BIS- and CogAT-battery and (b) the relative low cross-battery correlations of the figural factors in the NF-model because reliable figural variance was already bonded to a large extent with the considerably figural-based *g*-factors. One reviewer pointed out that *g* tended often to be the predominant source of variance in indicators as compared to specific abilities (cf. Canivez & Kush, 2013), although this interpretation was often based on HO-models. In examining the standardized factor-loadings the pattern of the indicators (Table 3) in our NF-model showed no such general predominance of the general factor, especially for the verbal (and for the numerical) ability-factor. This relative *g*-dominance of the figural (vs. numerical and verbal) subtests corresponds with the outlined pattern of less evidence for convergent–divergent validity of these figural group factors. At this point, we recommend caution in making a more substantial interpretation of the different convergent correlations of the content-facet-specific factors because the number of tests was limited (for pragmatic reasons) to two to five tests per test-battery-specific content-facet factor. It would be interesting to replicate these results with a broader range of tests. As shown in prior studies (e.g., Brunner et al., 2012; Reise, 2012), the model-based reliability estimates of the domain-specific factors in the NF-model as operationalized by omega hierarchical [ω_h] using factor loadings were rather low. The loadings we made use of for estimating omega hierarchical stemmed from the complex NF-model with all batteries and, thus, were partially rather low. Half of the variables in the NF model loaded substantially above .30 on both *g* and the group-factor, but nearly half of the loadings reached substantial values only on the general- or the respective group-factor in consequence of modeling orthogonal factors. As can be seen from the formula of ω_h, the reliability estimates for the specific ability factors are influenced substantially by the amount of the *g*-factor loadings of the corresponding subtests. It seems that this was particularly relevant for the figural content factors, because the *g*-loadings of these subtests were particularly high. In contrast, the reliability coefficients of the verbal factors were rather higher, because a substantial amount of systematic verbal variance (i.e., loadings on the verbal factor) was set against the relatively moderate *g*-factor loadings of the corresponding subtests. So the interpretation of the subscales as precise indicators of unique constructs seems to be limited (Reise, 2012, p. 691), at least at first glance. Additionally it should be kept in mind, that reliability represents merely *one* aspect when evaluating the quality of a psychometric measure besides important other criteria as, for example, the (convergent) validity with similar constructs and/or external criteria. Correspondingly, the substantial convergent nested factor correlations of the verbal and numerical factors can be interpreted as the lower bound of the reliabilities of these factors. (Similarly, the *g*-factor correlations were substantial and very high, despite model-based reliability estimates being rather moderate.) In sum, this convergent–divergent correlation pattern of the nested factors gave us reason to believe that these beside-*g* and content-facet-specific (nested) factors have substantial relevance (at least, for the verbal and numerical factors). One anonymous reviewer asked justifiably about the implications for the magnitudes of cross-battery correlations of the group factors. In several studies *g* is observed as the predominant source of systematic variance of cognitive tasks and the factors beside/below *g* are sometimes designated as rather negligible. The purpose of our study was to systematically examine this variance not accounted for by *g*; and the findings suggest that it is fruitful to have a closer look on these former classified "rather negligible" variance components. In a next step it would be interesting to investigate the correlations of these content-specific nested factors with real life criteria (e.g., achievement tests and grades) to further evidence criterion validity of the content-facet group-factors.

However, there are some limitations to our study: although partially replicating former analyses and results that were primarily based on adult samples, the presented results are restricted to German high school students and a specific grade (9th graders; although this relatively age-homogeneous sample reduced corresponding confounding variance of different age cohorts). The students' participation depended partially on decisions of their principals as well as their teachers and parents. Therefore, the selection of the students was non-random. The high participation rate within the sample prevented a substantial bias due to a systematic self-selection within the targeted sample. Focusing on potential variance-restriction[2] we inspected the correlation-coefficients of our sample in comparison to those of the German CogAT norm sample. Effect-size *q* ranged from .03 to .30 indicating (at most) small to medium effects (Median: $q = .16$). These results suggest, that

[2] An inspection of the corresponding quotients of the standard deviations revealed that most standard deviations of our sample seem (very) similar to those standard deviations of the (Gymnasium) norm samples — not indicating severe range restrictions (SPM: 0.9; CogAT: 1.0 [WS], 1.0 [WL], 0.8 [MV], 1.0 [ZR], 1.7 [FA], 1. [FK]). We furthermore inspected these quotients calculated with the standard deviations of the whole CogAT- and SPM-norm samples (not just students attending the Gymnasium): as expected, these values were a bit higher (SPM: 1.5; CogAT: 1.4 [WS], 1.1 [WL], 1.4 [MV], 1.5 [ZR], 1.9 [FA], 1.5 [FK]) — but exceeding 1.5 for only one test.

range restriction was indicated to a small to medium extent as would have been expected for a (*Gymnasium*) high-school-sample. Although variance-restriction affects the magnitude of correlations, the *g*-factor-correlations reached magnitudes as high as can be expected from other studies. Nevertheless, our results can be seen as conservative estimates. One can assume that similar analyses in more heterogeneous samples would likely yield a (more) stable correlation pattern, particularly for the specific-ability-factors. Similar investigations based on different samples from other countries should be conducted. The analyses were based on a relatively broad and heterogeneous sample of 26 different intelligence tests, but practical limitations made it impossible to further broaden the range of different cognitive tasks and tests. These practical reasons limited the test-battery-specific content-facet specific group factors to two (CogAT) to five (BIS) tests per factor. Nevertheless, the test selection was guided by considerations to conceptually replicate prior results (from the first two research questions). It would still be fruitful to include even more tests of an even higher variety in order to further strengthen the interpretations. Although the different hierarchical models were derived from prior analyses, the HO- and NF-models in particular were specified by splitting the corresponding CogAT-tests (in order not to risk under-identifications).

In sum the findings of our study affirm the main results: (1) in the framework of CFA-modeling of hierarchically structured intelligence models, the NF-model represents a useful and fruitful alternative to GF- and HO-models especially when the general factor and domain-specific factors are of interest (conceptually replicating, e.g., Brunner et al., 2012; Gignac, 2005, 2006b; Gustafsson & Balke, 1993). (2) The general-factor as an underlying and broad intelligence-factor correlated very high proof consistency (almost) regardless of the instrument, given that it is developed to measure *g* (conceptually replicating Johnson et al., 2004, 2008). (3) In addition to and beyond prior results, test-battery-specific and content-domain-specific verbal and numerical ability factors proved consistent across independently developed test-batteries. For this reason, these results confirmed empirically the comment made by Johnson et al. (2008, p. 91) that "[t]here are substantive correlations among … specific abilities from battery to battery … and different tests measure them with reliability comparable to that associated with the general factor", at least and especially for the verbal (and, to a lesser degree numerical) content-facet factors.

References

Beauducel, A., & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin model of intelligence structure (BIS). *European Journal of Psychological Assessment*, *18*, 97–112.

Bentler, P. M. (1995). *EQS: A structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Brunner, M., Gogol, K. M., Sonnleitner, P., Keller, U., Kraus, S., & Preckel, F. (2013). Gender differences in the mean level, variability, and profile shape of student achievement: Results from 41 countries. *Intelligence*, *41*, 378–395.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796–846.

Brunner, M., & Süss, H. -M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*, 227–240.

Brunner, M., & Süss, H. M. (2007). Wie genau können kognitive Fähigkeiten gemessen werden?: Die Unterscheidung von Gesamt- und Konstruktreliabilitäten in der Intelligenzdiagnostik für den Berliner Intelligenzstrukturtest [How precisely can cognitive abilities be measured? The distinction between composite and construct reliabilities in intelligence assessment exemplified with the Berlin Intelligence Structure Test]. *Diagnostica*, *53*, 184–193.

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

Canivez, G. L., & Kush, J. C. (2013). WAIS-IV and WISC-IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). *Journal of Psychoeducational Assessment*, *31*, 157–169.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.

Colom, R., Abad, F. J., Garcia, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's Full Scale IQ, and *g*. *Intelligence*, *30*, 449–462.

de Wolf, C. J., & Buiten, B. (1963). Een factorenanalyse van vier testbatterijen [A factor analysis of four test batteries]. *Nederlands tijdschrift voor de Psychologie*, *18*, 220–239.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–482.

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. G., et al. (1974). Near identity of cognitive structure in two ethnic groups. *Science*, *183*, 338–339.

Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence*, *37*, 453–465.

Gignac, G. E. (2005). Revisiting the factor structure of the WAIS-R: Insights through nested factor modeling. *Assessment*, *12*, 320–329.

Gignac, G. E. (2006a). A confirmatory examination of the factor structure of the Multidimensional Aptitude Battery (MAB): Contrasting oblique, higher-order, and nested factor models. *Educational and Psychological Measurement*, *66*, 136–145.

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*, 73–86.

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

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79–132.

Guilford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 225–266). New York: Wiley.

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

Hakstian, A. R., & Cattell, R. B. (1975). *The Comprehensive Ability Battery.* Champaign: Institute for Personality and Ability Testing.

Heller, K. A., Kratzmeier, H., & Lengfelder, A. (1998). *Matrizen-Test-Manual Band 1 zu den Standard Progressive Matrices von J. C. Raven [Test manual of the standard progressive matrices of J. C. Raven].* Göttingen, Germany: Beltz-Test.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision [Cognitive ability test for the 4th to the 12th grade, revised].* Göttingen, Germany: Beltz-Test.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Hu, L., & Bentler, P. M. (1998). Fit Indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453.

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

Jäger (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intelligence achievement: Experimentally controlled, further development of a descriptive intelligence structure model]. *Diagnostica*, *28*, 195–226.

Jäger, A. O., Süss, H. -M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur Test, Form 4.* Göttingen: Hogrefe.

Jensen, A. R. (1998). *The g-factor.* Westport, CN: Praeger.

Jensen, A. R., & Weng, L. -J. (1994). What is a good g? *Intelligence*, *18*, 231–258.

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

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence*, *36*, 81–95.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) 'general intelligence', objectively determined and measured. *Journal of Personality and Social Psychology*, *86*, 96–111.

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small single-factor models do not adequately represent *g*. *Intelligence*, *39*, 418–433.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362.

McDonald, R. P. (1999). *Test theory: A unified treatment.* 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*, 1–10.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus the higher-order models of human cognitive ability structure. *Intelligence*, *41*, 407–422.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*, 137–150.

Ree, M. J., & Earles, J. A. (1991). The stability of *g* across different methods of estimation. *Intelligence*, *15*, 271–278.

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

Schmiedek, F., & Li, S. -S. (2004). Toward an alternative representation for disentangling age-associated differences in general and specific cognitive abilities. *Psychology and Aging*, *19*, 40–56.

Schwarz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, *6*, 461–464.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Süss, H. M., Kersting, M., & Oberauer, K. (1991). Intelligenz und Wissen als Prädiktoren für Leistungen bei computersimulierten komplexen Problemen [Intelligence and knowledge as predictors of achievement in computer simulated complex problems]. *Diagnostica*, *37*, 334–352.

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*, 261–288.

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences*, *8*, 585–586.

Vernon, P. A. (1989). The generality of *g*. *Personality and Individual Differences*, *10*, 803–804.

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

Weiss, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2 [CFT 20-R. Cattell Culture Fair Test, Scale 2].* Göttingen, Germany: Hogrefe.

Yung, Y. -F., Thissen, D., & McLeod, L. D. (1999). On the relation between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *65*, 113–128.