Investigating the ‘g’-saturation of various stratum-two factors using automatic item generation

Martin E. Arendasy *, Andreas Hergovich, Markus Sommer

University of Vienna, Department of Psychology, Psychometric Technology Group, Austria

Received 9 December 2006; received in revised form 27 November 2007; accepted 27 November 2007
Available online 4 February 2008

Abstract

Even though researchers agree on the hierarchical structure of intelligence there is considerable disagreement on the g-saturation of the lower stratum-two factors. In this article it is argued that the mixed evidence in the research literature can be at least partially attributed to the construct representation of the individual tests used to measure the stratum-two factors. In the study described here two top-down approaches to automatic item generation were used to build construct representation directly into the items. This enabled a clearer substantive interpretation of the shared commonalities extracted by the various stratum-two factors and helped to rule out alternative, but mathematically equivalent, structural models based on independent empirical evidence. This approach was used to investigate the g-saturation of five stratum-two factors using a sample of 240 male and female respondents. The results support the assumption that fluid intelligence exhibits the highest g-saturation.

© 2007 Elsevier Inc. All rights reserved.

Keywords: CHC models; g-saturation; Construct representation; Automated item generation

1. Theoretical background

Despite a growing consensus on a hierarchical structure of human intelligence with g at the apex and several broader stratum-two factors, researchers still disagree on the g-saturation of the broader stratum-two factors.

Recently some researchers (cf. Gignac, 2006; Robinson, 1999) have argued that general intelligence is more closely associated with crystallized intelligence (G_c). This claim is based on the results of exploratory and confirmatory factor analyses of Wechsler-like test batteries. For instance, Gignac (2006) investigated the g-saturation of the subtests included in different Wechsler-like test batteries. The author used a single-trait-correlated uniqueness confirmatory factor analysis and specified a g-factor in addition to correlated residuals between subtests assumed to measure either (1) crystallized intelligence (G_c), (2) visual processing (G_v), or (3) fluid intelligence (G_f) in order to control for their shared commonalities when estimating their g-saturation. The results indicated that the crystallized intelligence (G_c) subtests had the highest average factor loadings on the g-factor. This is in contrast to the results obtained by Roberts, Goff, Anjoul, Kylonen, Pallier and Stankov (2000) as well as Gustafsson (1984; 2002; Gustafsson & Balke, 1993; Undheim & Gustafsson,
1987), who reported that fluid intelligence ($G_f$) was virtually indistinguishable from the higher-order g-factor in their studies.

The picture is further complicated by results obtained by Bickley, Keith and Wolfe (1995), who investigated the factorial structure of 16 subtests of the revised Woodcock–Johnson battery (WJ-R) using a hierarchical confirmatory factor analysis. These researchers reported an almost equal g-saturation of fluid intelligence ($G_f: .88$), quantitative reasoning ($G_q: .86$) and crystallized intelligence ($G_c: .87$). This suggests that all three stratum-two factors are equal with regard to their g-saturation. This conclusion was further supported by the finding that the model fit decreased significantly once the standardized factor loadings of the three stratum-two factors on ‘g’ were set to 1. These results were also replicated by Carroll (2003), who used a more extended set of variables taken from the WJ-R norm sample.

Taken together, the literature provides rather mixed evidence with regard to the g-factor saturation of the stratum-two factors fluid intelligence ($G_f$), crystallized intelligence ($G_c$) and quantitative reasoning ($G_q$). In the following section we discuss several reasons for this mixed evidence.

2. Reasons for the mixed evidence on the g-factor saturation

The inconsistencies in the results on the g-saturation of the stratum-two factors can be attributed to differences between the studies in: (1) the stratum-two factors and (2) the homogeneity of the sample investigated, as well as (3) the construct representation (Embretson, 1983) of the individual subtests. In the following discussion we focus on the latter reason and outline how it affects results obtained with confirmatory factor analyses.

In several studies (e.g. Flanagan & McGrew, 1998; Johnson & Bouchard, 2005; McGrew, 1997) some subtests have been observed to load on more than one stratum-two factor. This has been taken as evidence that some subtests are far from pure measures of the intended stratum-two factor and would be best described as mixed measures. This claim is in line with research on the cognitive processes respondents use to solve a given item. This problem is commonly referred to as the experts’ blind spot (cf. Nathan & Petrosino, 2003). This problem is highlighted in Ashton and Lee’s (2006) criticism of the model specifications chosen by Gignac (2006). Using the same data set and method of analysis, the authors specified alternative correlated uniqueness based on their own task analysis and demonstrated that their model not only fitted the data equally well but was in fact psychometrically identical to the model specified by Gignac (2006). However, the model specified by Ashton and Lee (2006) led to substantively different results with regard to the g-saturation of the individual subtests. Their results indicated that the non-crystallized subtests had a higher average factor loading on the g-factor than the crystallized subtests.

Alternatively one could circumvent the problem associated with the construct representation (Embretson, 1983) of the individual subtests by building the construct representation directly into the items. However, this requires a systematic theory-driven item construction approach that seeks to maximize the construct-related variance in the item difficulties and simultaneously minimizes unwanted variance arising from non-construct-related factors. In the following section we will thus outline such a theory-driven item construction approach using the current debate on the g-saturation of the broader stratum-two factors of the Cattell–Horn–Carroll model (CHC model) outlined above. However, it should be noted that the approach could also be applied to research on alternative models of intelligence in order to circumvent problems arising from issues of construct representation within the framework of these models.
3. Automatic Item Generation as a tool for building construct representation into the item construction process

In the last few years Automatic Item Generation (AIG) has been introduced as a new item construction technology in applied psychometrics (Irvine, 2002). According to the various top-down approaches to automatic item generation, the construction process of the item generator starts with the definition of the latent trait to be measured. In the present study we used global definitions of the stratum-one factor as our starting point (cf. McGrew, 1997). In a second step a literature review was conducted to identify the cognitive processes, solution strategies and knowledge structures that characterize the latent trait. This literature review took account of research on individual differences as well as studies in cognitive psychology in order to identify item features that affect the cognitive processes and solution strategies respondents use to solve the given items. These item features constitute the radicals (Irvine, 2002) that are incorporated into the generative framework of an automatic item generator (cf. Arendasy, 2005; Arendasy & Sommer, 2005, 2007). These radicals are assumed to systematically affect certain properties of an item — e.g. its difficulty — due to their effect on the cognitive processes used by respondents to solve the items.

Arendasy et al. (Arendasy, 2005; Arendasy & Sommer, 2005, 2007) have recently pointed out that the specification of radicals (Irvine, 2002) alone does not suffice to ensure (1) a high psychometric quality (e.g. fit of the Rasch Model: Rasch, 1960/1980) and (2) a satisfying level of construct representation (Embretson, 1983). The authors argue that items must be generated in a manner that triggers solution strategies inherently related to the latent trait to be measured and simultaneously reduces interfering variance arising from non-construct-related cognitive processes. This latter aim is accomplished through the use of a set of ‘functional constraints’ that must be integrated into the quality control framework (cf. Arendasy, 2005; Arendasy & Sommer, 2005, 2007) of the item generator. The ‘functional constraints’ are derived from studies in cognitive psychology and psychometric experiments specifically designed to investigate their effect on the psychometric quality of automatically generated items (e.g. Arendasy & Sommer, 2005; Arendasy, Sommer, Gittler & Hergovich, 2006).

Once the item generator has been constructed and the necessity and sufficiency of the ‘functional constraints’ has been demonstrated in a series of psychometric experiments, the validity of the model can be evaluated in terms of its ability to predict the difficulties of the items generated on the basis of the specified radicals, using either multiple regression analyses or the linear logistic test model (LLTM: Fischer, 1973). If the radicals (Irvine, 2002) derived from the theoretical model contribute at least $R^2 = .80$ to the item parameters, the construct representation (Embretson, 1983) of the subtest can be assumed.

In this sense top-down approaches such as the automatic min-max approach (Arendasy, 2005; Arendasy & Sommer, 2007) outlined above serve to build construct representation directly into the item construction process by combining research in cognitive psychology, individual differences and applied psychometrics.

4. Formulation of the problem

Based on these theoretical and methodological considerations the aim of the present study was the investigation of the g-saturation of the stratum-two factors fluid intelligence ($G_f$), crystallized intelligence ($G_c$), quantitative reasoning ($G_q$) and visual processing ($G_v$) using a test battery consisting of ten subtests constructed by means of top-down approaches to automatic item generation in order to demonstrate the utility of this approach in research on the structure of human intelligence. In this test battery each of the five stratum-two factors is represented by two subtests. We chose this approach in order to ensure the construct representation (Embretson, 1983) of the individual subtests and allow for a less ambiguous model specification and interpretation of the commonalities captured by each of the five stratum-two factors.

5. Method

5.1. Procedure

The study was carried out during a university seminar held by the first author. Eight of the ten subtests were administered via TestWeb, which enables web-based psychological assessment. The remaining two subtests were administered using the Vienna Test System, which enables computerized psychological assessment. Participants were tested in groups of one to three respondents. The tests were presented in one session lasting around 2.5 h, including two breaks of about 10 min each. Each group of respondents was tested by a student test administrator, who was personally present during that time. The respondents were informed that they were going to work on different ability tests in three consecutive sessions and that there would be two breaks between the sessions. All respondents worked on the test battery in the same sequence: (1) Visual Short-term Memory, (2) Verbal Short-term Memory, (3) Arithmetic...
Table 1
Item generators, name of the test, task description and size of the item pool, radicals used in generative framework and their joint explanatory power regarding item difficulty parameters, psychometric benchmarks used to evaluate the potential of the radicals and functional constraints, references to prior studies

<table>
<thead>
<tr>
<th>Item generator</th>
<th>Name of the test</th>
<th>Task description</th>
<th>( K^a )</th>
<th>( N^b )</th>
<th>Radicals (generative framework)</th>
<th>( R^2 c )</th>
<th>Psychometric benchmark (quality control framework)(^d)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>EstiGen</td>
<td>CE: Computational Estimation</td>
<td>The goal of the task is to estimate the result of complex arithmetic problems.</td>
<td>80</td>
<td>746</td>
<td>• number sense • required level of estimation accuracy • arithmetic complexity • rounding difficulty</td>
<td>.88</td>
<td>RM / LLTM</td>
<td>Arendasy and Sommer (2007)</td>
</tr>
<tr>
<td>NGen</td>
<td>ANF: Arithmetic Fluency</td>
<td>The goal of the task is to find the proper arithmetic operations to complete an equation.</td>
<td>80</td>
<td>890</td>
<td>• number of arithmetic operators • number of different operators • arithmetic proficiency • identify ability of arithmetic operations</td>
<td>.90</td>
<td>RM / LLTM</td>
<td>Arendasy and Sommer (2007)</td>
</tr>
<tr>
<td>GeomGen</td>
<td>GEOM: Figural Inductive Reasoning</td>
<td>The goal of the task is to find the rules governing the figural matrices and to complete the matrices by applying these rules.</td>
<td>180</td>
<td>6817</td>
<td>• number of rules • rule complexity • number of elements • degree of abstraction</td>
<td>.92</td>
<td>RM / LLTM</td>
<td>Arendasy (2005), Arendasy and Sommer (2005)</td>
</tr>
<tr>
<td>ZafoGen</td>
<td>NID: Numerical Inductive Reasoning</td>
<td>The goal of the task is to find the rules governing the number series and to continue the series by applying these rules.</td>
<td>80</td>
<td>646</td>
<td>• length of the number series • rule complexity • number of rules • periodicity</td>
<td>.88</td>
<td>RM / LLTM</td>
<td>Arendasy, Hornke, Sommer, Häusler, Wagner-Menghin, Gittler, Bogner, and Wenzl (2007)</td>
</tr>
<tr>
<td>WGen</td>
<td>SYN: Synonyms</td>
<td>The goal of this task is to choose a German word from a list of four answer alternatives that has the same meaning as the given source word.</td>
<td>77</td>
<td>672</td>
<td>• word frequency of the source word • word frequency of the target word • word frequency of the distractors • semantic similarity of the source word and the answer alternatives</td>
<td>.89</td>
<td>RM / MRA</td>
<td>Arendasy et al. (2007)</td>
</tr>
<tr>
<td>WFGen</td>
<td>VF: Verbal Fluency</td>
<td>The goal of this task is to ensample character strings into a German word.</td>
<td>76</td>
<td>553</td>
<td>• word frequency • number of characters to be ensampled • relative amount of common character combinations • dissimilarity between the character string and the target word</td>
<td>.88</td>
<td>RM / MRA</td>
<td>Arendasy et al. (2007)</td>
</tr>
<tr>
<td>–</td>
<td>VI: Visual Short-term-Memory</td>
<td>The goal of the task is to remember and recall the location of icons presented on a schematic city map.</td>
<td>114</td>
<td>560</td>
<td>• number of icons to be remembered • size of the area enclosed by a circumference around the icons poor vs. good gestalt of the icon pattern</td>
<td>.88</td>
<td>RM / MRA</td>
<td>Hornke (2002)</td>
</tr>
</tbody>
</table>

(continued on next page)
Fluency, (4) Synonyms, (5) Figural Inductive Reasoning, (6) Computational Estimation, (7) Verbal Fluency, (8) Endless Loops, (9) Numerical Inductive Reasoning and (10) Spatial Comprehension. All the subtests were presented as computerized adaptive tests (CAT).

5.2. Measures

The test battery used in this study consisted of ten subtests. Two subtests were used to measure each of the five stratum-two factors quantitative reasoning (Gq), fluid intelligence (Gf), crystallized intelligence (Gc), short-term memory (Gsar) and visual processing (Gv). The selection of the subtests was based on the following criteria: (1) the item construction process of each subtest should be based on a theoretical model specifying the cognitive processes needed to solve the items, (2) the items should have been constructed using a top-down approach to automatic item generation, (3) in prior studies the radicals contributed at least $R^2 = .80$ to the item difficulty parameters and (4) the radicals of subtests assumed to load on different factors should not overlap (for details see Table 1).

In using these criteria we ensured that the construct representation (Embretson, 1983) of each subtest could be assumed based on independent psychometric evidence obtained in prior validation studies by means of multiple regression analyses and/or the linear logistic test model (Fischer, 1973).

5.3. Sample

The sample consisted of 126 (52.5%) male and 114 (47.5%) female respondents aged between 16 to 63 years ($M=37.11; SD=12.13$). The median age was 36 years. A total of 25 (10.4%) respondents had completed 9 years of schooling but no vocational training, while 91 (37.9%) respondents had completed vocational education. Ninety-nine (41.3%) respondents had a high school leaving certificate qualifying them for university entrance, and 25 (10.4%) respondents had graduated from university or college. Neither the gender distribution ($\chi^2 [1] = .04; p = .831$) nor the age distribution ($\chi^2 [9] = 16.45; p = .058$) differs significantly from that expected on the basis of the Austrian census of 2001. Furthermore, the distribution of educational level ($\chi^2 [4] = 9.44; p = .051$) was also in line with expectations based on the Austrian census of 2001. The sample can thus be regarded as representative within the age range investigated.

<table>
<thead>
<tr>
<th>Item generator</th>
<th>Name of the test</th>
<th>Task description</th>
<th>$K^a$</th>
<th>$N^b$</th>
<th>Radicals (generative framework)</th>
<th>$R^2^c$</th>
<th>Psychometric benchmark (quality control framework)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE: Verbal Short-term-Memory</td>
<td>The goal of the task is to remember and recall station names that are presented sequentially along a bus route shown on a schematic city map.</td>
<td>114</td>
<td>560</td>
<td>• number of bus stops to be remembered • presentation rate of the individual bus stops • level of concreteness of the bus stop names</td>
<td>.88</td>
<td>RM / MRA</td>
<td>Hornke (2002)</td>
<td></td>
</tr>
<tr>
<td>3DC: Spatial Comprehension</td>
<td>The goal of this task is to mentally rotate the distracter cubes and compare them to the target cube.</td>
<td>142</td>
<td>1298</td>
<td>• number of rotations • pattern complexity • position of the correct answer alternative • similarity of the distractors</td>
<td>.80</td>
<td>RM / LLTM</td>
<td>Gittler (1990)</td>
<td></td>
</tr>
<tr>
<td>EsGen</td>
<td>ELT: Endless Loops</td>
<td>The goal of this task is to determine from which side a picture of a given object (endless loop) has been taken.</td>
<td>160</td>
<td>4196</td>
<td>• visual complexity of the loops • 3D-complexity • global dominance effect • complete overlap of loop elements</td>
<td>.88</td>
<td>RM / LLTM</td>
<td>Arendasy (2000, 2005), Gittler and Arendasy (2003)</td>
</tr>
</tbody>
</table>

$a \ k$=number of items in item pool.  
$b \ N$=size of calibration sample.  
$c \ R^2$=proportion of variance of item difficulty parameters explained by radicals.  
Based on the CHC model, we assumed that the observed correlations between these ten subtests can be explained by five stratum-two factors and a higher-order g-factor. According to this theoretical model, the subtests ‘Arithmetic Flexibility’ and ‘Computational Estimation’ would load on $G_M$, while the subtests ‘Figural Inductive Reasoning’ and ‘Numerical Inductive Reasoning’ were assumed to load on $G_F$. ‘Synonyms’ and ‘Verbal Fluency’ were assumed to load on $G_c$, while the subtests ‘Verbal Short-term Memory’ and ‘Visual Short-term Memory’ would load on $G_{sar}$. The subtests ‘Endless Loops’ and ‘Spatial Comprehension’ were assumed to load on $G_v$. Furthermore, the five stratum-two factors were assumed to load on a higher-order g-factor (cf. Bickley et al., 1995; Carroll, 1993, 2003). The variance of the higher-order g-factor was set to 1 for identification and scaling purposes. In the following sections this model is referred to as the CHC model.

The CHC model was contrasted with two alternative models. In the first alternative model we assumed that the correlations between the ten subtests can be explained by a single g-factor. This model is thus referred to as the pure g-factor model.

In the second alternative model we assumed that the correlations between the ten subtests can be explained by three stratum-two method factors and a higher-order g-factor. In this model we assumed that the subtests ‘Arithmetic Flexibility’, ‘Computational Estimation’ and ‘Numerical Inductive Reasoning’ load on a numerical intelligence factor ($N$), while the subtests ‘Synonyms’, ‘Verbal Fluency’ and ‘Verbal Short-term Memory’ load on a verbal intelligence factor ($V$), and the subtests ‘Endless Loops’, ‘Spatial Comprehension’, ‘Figural Inductive Reasoning’ and ‘Visual Short-term Memory’ load on a figural intelligence factor ($F$). Furthermore, these three content factors were assumed to load on a higher-order g-factor. Even though this model is not supported by prior research on the construct representation of our subtests, we specified this model to rule out the possibility that the correlations between our ten subtests can be explained equally well by an alternative model. This model is referred to as the content factor model.

5.4. Tested models
assumption that the standardized factor loading of $G_q$ on the higher-order g-factor can be set to 1 ($G = G_q$ model), while the same assumption has been made with regard to the standardized factor loading of $G_c$ on the higher-order g-factor in the third restricted CHC model ($G = G_c$ model). The fourth model tested the assumption that $G_v$ is virtually indistinguishable from the higher-order g-factor by setting the standardized factor loading of $G_v$ on the g-factor to 1 ($G = G_v$ model). Finally we tested the assumption that the stratum-two factors $G_f$, $G_c$, $G_q$ and $G_v$ do not differ in their g-saturation by constraining the factor loadings of $G_f$, $G_c$, $G_q$ and $G_v$ on the higher-order g-factor equal. This model is referred to in the following sections as the $G \rightarrow G_f = G_c = G_q = G_v$ model.

6. Results

The data were analysed using AMOS 5.0 (Arbuckle, 2003). The results are presented in two sections. The first section presents descriptive statistics for the individual subtests. The second section describes the results of confirmatory factor analyses conducted using Maximum Likelihood estimation to calculate the parameters of the postulated models.

The means, standard deviations and reliability estimates of the subtests are presented in Table 2. Note that all measures meet the standard criteria for univariate normality as indicated by the fact that the skew and kurtosis values for the individual measures are $\leq 2$ (cf. Kline, 1998). In addition, the assumption of multivariate normality can be retained. The global fit of the models was assessed using the following cut-off values for the global fit indices (cf. Browne & Cudeck, 1993; Byrne, 2001; Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004): non-significant $\chi^2$-test, $\chi^2/df < 2$, RMSEA $\leq .06$, SRMR $\leq .06$ and CFI $\geq .95$. The fit statistics of all the models tested are presented in Table 2 (below).

As Table 2 shows, the CHC model provides a good fit to the data. Only the $\chi^2$-test turned out to be significant. However, this can be partially attributed to the sample size used in this study (cf. Byrne, 2001). An inspection of the standardized residuals further revealed that none of them was statistically significant at $\alpha = .05$; this provides further evidence of the fit of the CHC model to our data.

In case of the g-factor model, none of the fit indices meets the standard criteria for an adequate model fit. The superiority of the CHC model is confirmed by the improvement in the $\chi^2$ statistic over the pure g-factor model ($\Delta \chi^2 [4] = 46.55; p < .001$). Furthermore, the g-factor model also fits significantly less well ($\Delta \chi^2 [3] = 35.50; p < .001$) than the content factor model. However, an inspection of the standardized factor loadings of the ten subtests on their corresponding content factors revealed that their magnitude was generally low and in several cases even non-significant ($p > .05$). On the basis of this result the content factor model was rejected.

In a further step we examined the statistical significance of the factor loadings of the CHC model, since it provided the best model for our data. The factor loadings of the individual subtests on their corresponding
stratum-two factors were all moderate to high and reached statistical significance at $\alpha = .01$. The standardized factor loadings are depicted in Fig. 1.

The results also indicated that the g-factor is nearly indistinguishable from $G_f$. The standardized factor loading of .98 is close to 1.00, and the error variance of $G_f$ did not reach the significance level ($p > .05$) and was thus excluded. It was thus hardly surprising that the $G=f$ model not only exhibited a good fit to the data but fitted the data no worse than the unrestricted CHC model ($\Delta \chi^2 [1] = 1.31; p = .250$).

With regard to the $G=G_q$ model, the fit statistics indicate a mediocre model fit and a $\Delta \chi^2$ test ($\Delta \chi^2 [1] = 15.19; p < .001$) revealed that the $G=G_q$ model fits significantly worse than the unrestricted CHC model. Similar results were obtained for the $G=G_c$ model and the $G=G_c$ model. Both models failed to meet the standard criteria of an adequate or a good model fit and also fitted the data significantly worse than the unrestricted CHC model. The results of the model comparisons ($G= \rightarrow G=f=G_q=G_c$ model, the fit statistics and the $\Delta \chi^2$ test ($\Delta \chi^2 [2] = 112.85; p < .001$) contradict the assumption that the four stratum-two factors contribute equally to the higher-order g-factor.

Taken together the results obtained in this study support the position that $G_f$ is virtually indistinguishable from the g-factor, while this is not the case for the remaining stratum-two factors.

7. Discussion

Even though there is a consensus on the hierarchical structure of human intelligence, researchers still disagree on the g-saturation of various stratum-two factors due to mixed evidence in the literature (cf. Ashton & Lee, 2006; Bickley et al., 1995; Carroll, 1993, 2003; Gignac, 2006; Gustafsson, 1984, 2002; Gustafsson & Balke, 1993; Roberts et al., 2000; Robinson, 1999; Undheim & Gustafsson, 1987). In this article we have argued that the mixed evidence can be attributed to (1) differences in the stratum-two factors covered in the test batteries investigated and (2) the construct representation (Embretson, 1983) of the individual subtests. In previous studies several subtests have turned out to be mixed measures of two or more stratum-two factors, which cause problems in the substantive interpretation of these factors. We have argued that these problems are closely linked to the construct representation (Embretson, 1983) of the subtests. Since a confirmatory factor analysis does not establish the meaning of the extracted factors, some sort of independent evidence is needed in order to come up with a substantive interpretation of the shared variance captured by each stratum-two factor.

In this article we have suggested that top-down approaches to automatic item generation, such as the automatic min-max approach (cf. Arendasy, 2005; Arendasy & Sommer, 2005, 2007), can be used to provide this kind of information. In our own study we used results of prior studies regarding the construct representation (Embretson, 1983) of the individual subtests to select two subtests to measure each of five of the eleven stratum-two factors of the Cattell–Horn–Carroll model. Even though the application of the top-down approach to automatic item generation presented in this article is by no means restricted to the Carroll–Horn–Cattell model we chose this particular model of human intelligence due to the disagreement on the g-saturation of the broader stratum-two factors outlined in the current research literature in order to demonstrate the usefulness of our approach. With this aim in mind we resorted to prior studies regarding the construct representation (Embretson, 1983) of the individual subtests of the test battery used in this study to select two subtests to measure each of five of the eleven stratum-two factors of the Cattell–Horn–Carroll model. In order to investigate whether the Cattell–Horn–Carroll model indeed provides the best description of our data we also investigated the fit of a pure g-factor model and a content factor model. The results indicate that the Cattell–Horn–Carroll model provides a good fit to the data while both alternative models fail to meet the standard criteria for an acceptable model fit. It should be noted that we were unable to test all plausible alternative models of human intelligence such as the one proposed by Vernon (1964) and associates due to the lack of specific subtests in our test battery associated with these models. The conclusion that the Cattell–Horn–Carroll model provides the best fit is thus tentative. Given that the Carroll–Horn–Cattell model indeed provides the best fit to the data the comparison of different restricted CHC models suggests that $G_f$ is virtually identical to the higher-order g-factors. However, the identity of the $G_f$-factor and psychometric ‘g’ is sometimes taken as evidence of a misspecification of the Cattell–Horn–Carroll model (cf. Johnson & Bouchard, 2005). Proponents of this view argue that the correlations between the various subtests might be better explained by an alternative model. The results of the model comparisons carried out in our study indicate that this explanation can be ruled out at least for the two other plausible models we compared against the CHC model in this study.

The finding that $G_f$ and psychometric ‘g’ are virtually identical nevertheless poses the question of whether $G_f$ is misclassified as a stratum-two factor and instead represents the essence of psychometric g. This question is
particularly interesting in the light of current revisions of Cattell’s investment theory. Cattell (1963) originally assumed that people utilize their fluid intelligence in order to acquire proficiency in various other mental abilities. However, the limited empirical examinations of the investment theory failed to provide strong evidence for a directional relation between $G_f$ and $G_c$ (cf. Blair, 2006). Moreover, these studies indicated that $G_f$ and $G_c$ are not only distinct early in the lifespan but their developmental course is also different (cf. Horn & Noll, 1997; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Thus more recent revisions of the investment theory assume that the influence of $G_f$ on $G_c$ is mediated by learning (Schweizer & Koch, 2002) and that fluid intelligence and learning share a common cognitive basis—namely capacity or executive functioning and processing speed. This revision turned out to be rather successful in predicting knowledge gain. Also note that the theoretical model bears some similarities to the one outlined by Anderson (2001). The theoretical model proposed by Schweizer and Koch (2002) is particularly interesting with regard to the radicals (Irvine, 2002) of our two measures of fluid intelligence. The radicals ‘number of elements / length of the number series’ and ‘number of rules’ are commonly assumed to represent aspects of capacity, mental load and executive functioning while the radical ‘rule complexity’ is assumed to be linked to processing speed (for an overview: Arendasy & Sommer, 2005). Furthermore, expert-novice research (cf. Bransford, Brown & Cocking, 2000) has indicated that respondents who scored high on fluid intelligence measures posed more questions while learning various subject matter and also exhibited a higher knowledge gain. This might point to the influence of abstraction ability, which constitutes the third set of radicals (Irvine, 2002) in our fluid intelligence measures. One would thus expect $G_f$ to exhibit the highest level of $g$-saturation in cross-sectional studies of the structure of human intelligence. Results such as those presented by Gignac (2006) and Robinson (1999) are therefore not in line with these theoretical assumptions.

Furthermore, the approach we used in this study also has some interesting implications for previous studies. Even though automatic item generation (AIG) methods can help to solve several problems which arise with regard to (1) the substantive interpretation of the extracted factors and (2) the choice between competing models we do not intend to state that one approach (CFA models using tests constructed by AIG) should entirely replace the other one (CFA analyses of traditional test batteries) since both have their own strength and weakness. In fact the two approaches do not necessarily exclude each other. For instance one could use methods of psychometrically modelling of the item difficulties in traditional subtests to provide independent evidence that can be used to distinguish between psychometric identical but substantially different CFA models. Such an analysis would—for instance—be useful in cases of the models Gignac (2006) and Ashton and Lee (2006) proposed. Theory based approaches to automatic item generation such as the cognitive design system approach (cf. Embretson & Gorin, 2001) and the automatic min-max approach (cf. Arendasy, 2005) even recognize this possibility as a means to derive a set of radicals in constructing an automatic item generator or in deciding between psychometrically identical but substantively different models on the structure of human intelligence.

References


Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In J. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 73–95). Mahwah: Erlbaum.


Roberts, R. D., Goff, G. N., Anjoule, F., Kylloinen, P. C., Pallier, G., & Stankov, L. (2000). The Armed Services Vocational Aptitude Battery (ASVAB)—Little more than acculturated learning (Gc)¿ Learning and Individual Differences, 12, 81–103.


