Psychological Test and Assessment Modeling, 2024.66:3-29

DOI: https://doi.org/10.2440/001-0010

Disentangling fluid and crystallized intelligence by means of Bayesian structural equation modeling and correlation-preserving mean plausible values

André Beauducel[1], *Richard Bruntsch*[2], *Martin Kersting*[3]

[1] University of Bonn, Department of Psychology, Bonn, Germany;
[2] GATEWAY Solutions AG, Assessment, Bern, Switzerland;
[3] Justus Liebig University of Gießen, Gießen, Germany;

Abstract:
The present study presents Bayesian confirmatory factor analyses of data from an extensive computer intelligence test battery used in the applied field of assessment in Switzerland. Bayesian confirmatory factor analysis allows to constrain the variability and distribution of model parameters according to theoretical expectations using priors. Posterior distributions of the model parameters are then obtained by means of a Bayesian estimation procedure. A large sample of 4,677 participants completed the test battery comprising 21 different tasks. Factors for crystallized intelligence, fluid intelligence, memory, and basic skills/clerical speed were obtained. The latter factor is different from speed-factors in several other tests as it encompasses speeded performance on moderately complex tasks. Three types of models were compared: for one type, only the expected salient loadings were freely estimated, and all cross-loadings were fixed to zero (i.e., independent clusters) whereas for the other two types of models normally distributed priors with a zero mean were defined. The latter two types were again altered regarding the amount of defined prior variance. Results show that defining substantial prior variances for the cross-loadings in Bayesian confirmatory factor analysis allow to overcome limitations of the independent clusters model. In order to estimate individual scores for the factors, mean plausible values were computed. However, the inter-correlations of the mean plausible-values substantially overestimated the true correlations of the factors. To improve discriminant validity of individual score estimates, it was therefore proposed to compute correlation-preserving mean plausible values. The findings can be applied to derive estimates for factorial scoring of a test battery, especially if cross loadings of subtests must be expected.

Correspondence:

André Beauducel, University of Bonn, Department of Psychology, Kaiser-Karl-Ring 9, 53111 Bonn, Germany, email: beauducel@uni-bonn.de

Keywords: Fluid intelligence; crystallized intelligence; Bayesian confirmatory factor analysis, mean plausible values

Introduction

Tests for the assessment of cognitive abilities (i.e., intelligence) are widely used in the applied field, for example, in the context of personnel selection, for clarifying assessments in the setting of clinical psychology and school psychology, and for the large-scale investigation of academic achievement. While there are short tests for the assessment of general mental ability, larger test batteries allow for the distinction of several of the broad dimensions of intelligence. Since Horn and Cattell (1966) and Carroll (1993), the Cattell-Horn-Carroll (CHC; McGrew, 2009) theory is often used as a frame of reference, although other models of intelligence are sometimes also discussed (Guilford, 1988; Guttman & Levy, 1991; Süß & Beauducel, 2015). Fluid intelligence/fluid reasoning (gf) and crystallized intelligence/comprehension-knowledge (gc) are among the most prominent general factors of the CHC model, although a clear representation of these factors also depends on methodological specifications (Carroll, 1995; Lang, Kersting, & Beauducel, 2016). This is not surprising as –since the early work of Spearman (1904)– there was a close relationship between different forms of multivariate analysis and the resulting models of intelligence.

Besides the effect of methodological specifications and multivariate analysis on models of intelligence, there is also an effect of task sampling on model results. A model of systematic sampling and classification of intelligence measures has been proposed by Guttman and Levy (1991). However, task sampling also occurs in applied settings where the fit of the tasks to demands of the job or a specific degree program is important. In these contexts, typically test development does not follow a systematic combination of task processes and content. Rather, when tests are developed in the applied field, task sampling is commonly based on the demands of the setting of its application. For example, the Scholastic Aptitude Test (SAT) has a clear focus on school achievement. This perspective was so dominant that the conceptual relationship between the SAT and cognitive abilities was neglected by test developers. Although cognitive abilities may be seen as a foundation of academic achievement (accordingly, a substantial empirical relationship of the SAT with intelligence has been noted; Frey, 2019; Frey & Detterman, 2004), the SAT was not developed for the assessment of intelligence and its tasks were sampled to the demands of the academic setting. In contrast, other tests developed for the applied field, like the Woodcock-Johnson IV Full Test Battery (Dombrowski, McGill, & Canivez, 2018) were explicitly related to a model of intelligence. Relating a test battery for the assessment of cognitive abilities not only to the demands from applied settings but also to the dimensions from an intelligence model allows for an improved understanding of the meaning and relevance of the test scores. In addition, the theoretical classification of the tasks also contributes significantly to validation. If test performance can be shown to represent

indicators of established ability constructs, the empirical evidence obtained in other studies for the validity of these constructs can be used as an argument for the test in question (i.e., in terms of validity generalization). Construct-based validity generalization may allow to relate the constructs measured by a given test battery to empirical findings on criterion validity of intelligence that have been obtained, for example, in the meta-analyses by Lang, Kersting, Hülsheger and Lang (2010), Salgado et al. (2003), Schmidt and Hunter (1998) and Sackett, Zhang, Berry and Lievens, (2021). Even if data on the test battery itself were not part of a corresponding meta-analysis, the test might be classified into the nomological network of abilities due to the construct explanation. This would allow for a construct-based validity generalization as it is encouraged by the DIN 33430 (DIN, 2016), a German quality standard.

By scoring a test battery developed according to practical demands in terms of established dimensions of intelligence, theory-based expectations may be derived for the test scores and a systematic investigation of construct validity of the test may be fostered. To explore the possibility to score a test battery from an applied setting according to the state-of-the-art of intelligence research (i.e., by means of factorial scoring), the present analyses were based on a Multicheck® test battery for economy/business and administration. The tasks of the Multicheck® test battery were gained on the basis of requirement analyses. Practitioners were asked which skills are important for training and work. The test items are intended to simulate real-life requirements, which fosters a high degree of face validity. For the following analyses, we reviewed these tasks, which were initially developed to represent practical requirements, and hypothesized which established constructs these tasks could be assigned to. We assume that the tasks can be attributed to the constructs (1) gf, (2) gc, (3) basic skills/clerical speed and (4) memory. Note that Schmitz and Wilhelm (2019) convincingly argued that clerical speed has meanwhile the status of a broad ability factor in hierarchical models of intelligence (see also Carroll, 1993). In order to emphasize the relevance of the factor for routine work in the office and administration, we use the term ‘basic skills/clerical speed’ in the following.

However, when tasks are developed for applied settings, they will not necessarily represent a pronounced simple structure, with single substantial loadings of each task on only one factor. It is likely that several cross-loadings occur when factor analyses of such task batteries are performed. In this way, large task batteries pose a challenge for factor analysis because the models have to follow theoretical expectations while the variables are expected to deviate from simple structure. Although confirmatory factor analysis would allow for a specification of loadings according to theoretical expectations, the multiple cross-loadings may result in extensive specification searches leading to inconsistent models and capitalization on chance (MacCallum, Roznowski, & Necowitz, 1992). To overcome this problem, Bayesian confirmatory factor analysis (BCFA) may be used instead of other forms of confirmatory factor analysis, as it allows for the specification of loadings according to theoretical expectations as well as the specification of the degree of variability of multiple cross-loadings. BCFA has been proposed in the context of Bayesian structural equation modeling (Muthén & Asparouhov, 2012), has been shown to identify population-loadings well (Xiao, Liu,

& Hau, 2019), and has meanwhile been applied in settings where complex loading patterns were expected (Weide, Scheuble, & Beauducel, 2021). A description of the technical details of BCFA in Mplus can be found in Asparouhov and Muthén (2010b). BCFA is based on a Markov Chain Monte Carlo (MCMC) algorithm and on splitting of the model parameters into groups (Gibbs sampler). Mplus defaults or optionally the user defines a prior distribution of parameters according to some expectations. If there is no clear expectation a flat distribution around zero is expected (non-informative prior). If a parameter is fixed (maximal specific expectation), an extremely small prior variance is used. The posterior distribution of each group of model parameters is generated from the conditional distribution of the remaining model parameters, the original data, and the prior distribution. At the end of the iterative sequence, posterior distributions of the parameters are constructed from the priors and the data according to the Bayes theorem.

While expected salient loadings will be freely estimated in BCFA, the size of non-salient loadings can be controlled by means of the prior variance when the mean prior is zero. A small prior variance will result in smaller cross-loadings and larger prior-variance will allow for larger cross-loadings (i.e., larger variation of the cross-loadings around zero). It has been recommended to try out models with different prior variance (Asparouhov, Muthén, & Morin, 2015). Hence, BCFA models without prior variance and models with different prior variances of cross-loadings will be compared in the present study. Comparing these models will give an account of the relevance of cross-loadings which could help to choose a final model that can be used for additional scoring of the Multicheck® test battery in the course of further development. As BCFA allows for the specification of theoretical expectations as well as multiple cross-loadings it is suitable for the investigation of large test batteries that were developed in applied settings and hence were not strictly designed to follow a given dimensional structure. However, as an alternative to the specification of non-zero prior variances for cross-loadings in BCFA, it would have been possible to specify inequality constraints for the cross-loadings. Inequality constraints, i.e., interval restrictions (e.g., LISREL, Jöreskog, & Sörbom, 2018) of cross-loadings (e.g., between -.30 and .30) can be specified in the context of maximum-likelihood estimation (Rindskopf, 2012). However, the cross-loadings cannot exceed the interval limits in this approach, which could be a problem, if there is no theoretical justification for a specific limit. A theoretical justification for a specific interval limit, e.g., why to use .30 instead of .35, could be impossible. Moreover, for parameters lying on the boundary of the parameter space of interval restrictions, the test statistic becomes a mixture of χ^2-distributions, whereas otherwise, it remains a χ^2-distribution (Savalei & Kolenikov, 2008). For these reasons, the BCFA approach based on restricted prior variances of cross-loadings is preferred over interval restrictions of cross-loadings in the present context. Therefore, the first aim of the present study is the investigation of theoretical expectations in a complex test battery that was developed in applied settings using BCFA.

Moreover, mean plausible values have been proposed as a method to compute scores for BCFA factors (Asparouhov & Muthén, 2010a). It has been noted that mean

plausible values and the best linear factor score predictor are nearly identical when the number of imputations and sample size is large (Beauducel & Hilger, 2022a). However, this implies that mean plausible values are not correlation-preserving, that is, that the inter-correlations of mean plausible values for BCFA factors are not the same as the inter-correlations of the BCFA factors themselves. This issue has already been discussed in the context of exploratory factor analysis and correlation-preserving factor scores have been proposed to overcome this issue (e.g., McDonald, 1981). Accordingly, it has been suggested to transform mean plausible values resulting from BCFA into correlation-preserving mean plausible values (Beauducel & Hilger, 2022b). Note that correlation-preserving mean plausible values have the same inter-correlations as the BCFA factors. As empirical research on this issue is lacking, the second aim of the present study is the investigation of the difference between the BCFA factor inter-correlations and the inter-correlations of the corresponding mean plausible values. It is of special interest to perform these analyses with data from intelligence tasks because a substantial amount of common variance and substantial factor inter-correlations can be expected in this context.

To sum up, BCFA is a suitable tool for the investigation of a theoretically expected factor structure in task settings with high factorial complexity as well as for the construction of scores for the intended factors in a specific test battery. As the possibility to define prior variances for cross-loadings is an essential difference between BCFA and other forms of confirmatory factor analysis, the first aim of the present study is a BCFA analysis of a version of the Multicheck® test battery by means of models without prior variance, with small prior variance, and with substantial prior variance. Thereby, it is possible to investigate whether BCFA with prior variance of cross-loadings allows for the identification of intelligence factors that have been established by previous research (e.g., Carroll, 1993) in a large test battery for the assessment of cognitive abilities used in applied settings. The second aim of the study is to compare the inter-correlations of the mean plausible-values with the inter-correlations of the BCFA factors. If the inter-correlations of the mean plausible values are substantially different from the inter-correlations of the BCFA factors, this indicates that correlation-preserving mean plausible values could be a valuable alternative.

Method

Participants

A sample of 4,677 German-speaking participants (Swiss residents; 2,340 females; age: $M = 15.63$, $SD = 2.54$ years) completed the Multicheck® computer-test battery (version "*Economy and Administration*") in order to get feedback on their aptitude for jobs in the context of vocational training in the sector of economy/business and administration (e.g., merchant).

Procedure

When they wish to apply for a training position (i.e., a job in the context of their vocational training), candidates for apprenticeships in Switzerland may be encouraged by the institution to submit their feedbacks from a Multicheck® assessment. The assessment was supervised and proctored in designated venues and took about 3.5 hours to complete. Participation was voluntary and participants could withdraw from participation at any time without any disadvantage, although an incomplete data set led to less meaningful (i.e., diminished) results in the individual feedback (as missing entries are set to zero in the scoring). The participants accept that their data are used for the requested feedback and for improvement of the test battery. Note that the participants were free to use the feedback for their job applications, but the assessment was not a part of a specific job selection procedure (as the test feedbacks belong to participants' property).

Measures

The 21 tasks of the test battery, their reliabilities in the total sample, and expected salient loadings are presented in Table 1. All included tasks were performance tasks in that correct and incorrect responses were possible. As an exception, the situational judgment tests for the assessment of social skills are scored gradually from 0 to 3 for each item. A short description of the tasks can be found in the Supplement. Cronbach's Alpha of the tasks was very diverse.

A classification of the tasks as marker variables for intelligence factors according to task content can be found in the right column of Table 1. Note that besides a rather clear classification of vocabulary- and knowledge-based tasks to gc, of reasoning-tasks to gf, and the classification of memory tasks to memory (M), a factor combining basic skills and clerical speed (BS) was expected. Tasks for the measurement of concentration and speeded choices of responses were expected to load on this factor. Moreover, social skills tasks were expected to load on this factor because they were based on situational judgement tests designed as simulations of practical situations at the workplace with rich demands to information processing and basic problem solving (Krumm et al., 2015).

Statistical Analysis

As the sample size is large enough, a random split of the sample according to odd and even case-numbers into two equal subsamples (subsample 1, $n = 2338$; subsample 2, $n = 2339$) was performed in order to investigate the robustness of the results. Age was similar in the subsamples (subsample 1: $M = 15.61$, $SD = 2.45$; subsample 2: $M = 15.64$, $SD = 2.64$; $t_{4675} = -0.46$, $p = .32$) as was the gender distribution (subsample 1: 49.4 % females; subsample 2: 50.7 % females). The BCFA analyses were performed

separately for each subsample with Mplus 8.4 (Muthén & Muthén, 2019). Factor variances were fixed to one and expected salient loadings were freely estimated. In Models 1 (for subsample 1) and 2 (for subsample 2) only the expected salient loadings were freely estimated and all cross-loadings were fixed to zero. According to Asparouhov and Muthén (2017), a fixed-to-zero loading can be treated as a loading with a prior mean of zero and a prior variance of zero. Moreover, the output does not provide parameter estimates for the fixed to zero loadings. This implies that the fixed zero loadings were not adapted to the data in Models 1 and 2.

Table 1. Test battery, their reliabilities, and expected salient loadings

	Cronbach´s Alpha	expected salient loading
grammar (German)	.87	gc
orthography (German)	.94	gc
text comprehension (German)	.83	gc
vocabulary (German)	.64	gc
computer knowledge[a]	.82	gc
grammar (English)	.86	gc
communication (English)	.87	gc
vocabulary (English)	.51	gc
figural analogies	.82	gf
verbal analogies	.43	gf
arithmetic	.78	gf
numerical estimation	.78	gf
relate information	.83	gf
figural memory	.60	M
verbal memory	.72	M
compare numbers	.96	BS
social skills – customer service	.34	BS
organizational skills	.96	BS
social skills – team	.40	BS
social skills – error handling	.26	BS
concentration[b]	-	BS

Note. BS = basic skills/clerical speed; M = memory; [a]the complete test also contains items of digital competences, but here, only knowledge items were used; [b]only the sum of correct responses was available; internal consistencies for the tasks were computed as standardized Cronbach Alpha coefficients with missing values set to zero in the item scoring; missing values occur when participants do not respond to an item during the test execution or when time runs out.

In the remaining models, non-salient loadings were estimated with normally distributed priors with a zero mean for gc, gf and BS. An overview of the parameter specifications in the different models is given in Table A1 (see Appendix). As there are only two variables for the measurement of M, this factor could only be estimated using an equality constraint on the respective expected salient loadings. To improve the robustness of the factor which is only based on two variables with expected salient loadings, non-salient loadings on M remained fixed to zero (no variability of cross-loadings on M was allowed). In each subsample two models with prior variance were performed, one with a prior variance of $\sigma^2 = 0.01$ (Models 3 and 4), one with $\sigma^2 = 0.05$ (Models 5 and 6). We followed the recommendation of Zitzmann and Hecht (2019) to use a potential scale reduction (PSR) smaller than 1.05. Therefore, we did not use the Mplus default for BCONVERGENCE of 0.05, but a BCONVERGENCE of 0.005 resulting in a PSR of 1.005. We ensured convergence by allowing 400,000 iterations when necessary. Mean plausible values were computed from 500 imputations, which according to Beauducel and Hilger (2022a) should be sufficiently large to approach optimal determinacy of mean plausible values. As the tasks, not the single items, were analyzed, all measured variables were specified as continuous variables. The further model specifications can be found in the Mplus input files example for Model 5 (see Appendix).

Starting from the idea that latent variables can be regarded as observed variables with missing values for all observations plausible values are computed like missing values (Asparouhov & Muthén, 2010a, 2022). The values are generated using MCMC simulation. For each plausible value, 100 MCMC iterations were performed which allows to obtain approximately the posterior distribution of the respective factor. In the present study, the mean across 500 plausible values was computed for each factor. The mean plausible values as well as their correlation-preserving version (Beauducel & Hilger, 2022b) were computed.

Results

The fit of the models in the two subsamples is given in Table 2. Obviously, the models without cross-loadings have a low comparative fit index (CFI), although two correlated errors were specified, one for *text comprehension (German)* and *verbal memory*, and another for *arithmetic* and *numerical estimation* (see Figure 1 for the conceptual diagram). As the fit of these models falls far below conventional criteria (Hu & Bentler, 1999), further improvements of model fit without allowing for non-zero cross-loadings would result in a substantial number of correlated errors.

Table 2. Model fit and *RMSD* of the completely standardized loadings.

							RMSD[b]			
Model	sub-sample	prior σ^2	*ppp*[a]	*BIC*	*RMSEA*	*CFI*	Model 3	Model 4	Model 5	Model 6
1	1	-	.000	239882.65	.065	.854	-	-	-	-
2	2	-	.000	238915.79	.067	.850	-	-	-	-
3	1	0.01	.000	239477.30	.052	.912	-	.035	.069	.072
4	2	0.01	.000	238522.62	.056	.903	.026	-	.070	.054
5	1	0.05	.000	239271.05	.048	.928	.056	.067	-	.042
6	2	0.05	.000	238285.73	.052	.921	.076	.073	.044	-

Note. [a]χ^2-based posterior predictor p-value, *BIC* = Bayesian Information Criterion; *RMSEA* = root mean square error of approximation; *CFI* = comparative fit index; *RMSD* = root mean square difference; [b]the *RMSD* for completely standardized loadings is given in the upper-triangle, and the *RMSD* for factor inter-correlations is given in the lower triangle.

Model fit was moderate but acceptable for the models allowing for cross-loadings (see Table 2). For the first subsample, models fit slightly better than for the second sub-sample. The Bayesian Information Criterium (*BIC*) is considerably smaller for the models with a prior σ^2 of 0.05 (Model 5 and 6), indicating that these models have an improved fit compared to the models based on a prior σ^2 of 0.01. As cross-loadings may result in some instability of the loading pattern, the root mean square difference (*RMSD*) between the completely standardized loadings of the models with non-zero prior variance was computed in order to evaluate the similarity of results across the two subsamples for each Model.

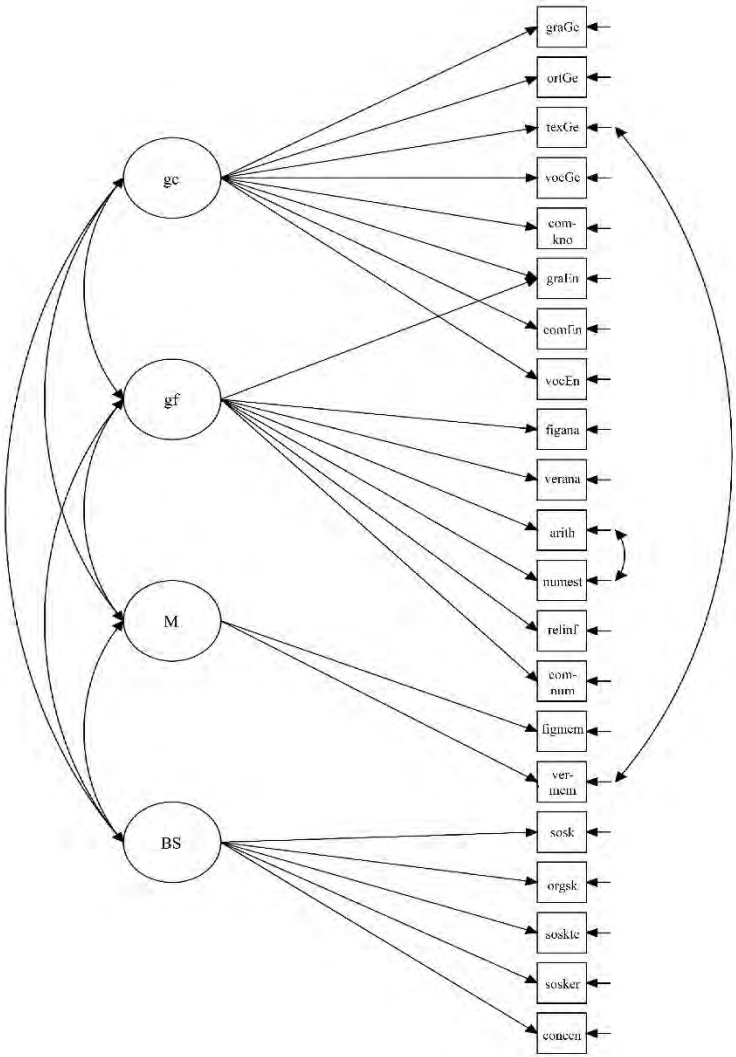

Figure 1. Conceptual diagram of BCFA freely estimated model parameters (without priors; graGe = grammar (German), ortGe = orthography (German), texGe = text comprehension (German), vocGe = vocabulary (German), comkno = computer knowledge, graEn = grammar (English), comEn = communication (English), vocEn = vocabulary (English), comnum = compare numbers, figana = figural analogies, verana = verbal analogies, arith = arithmetic, numest = numerical estimation, figmem = figural memory, vermem = verbal memory, sosk = social skills, orgsk = organizational skills – customer service, soskte = social skills – team, sosker = social skills – error handling, relinf = relate information, concen = concentration.

Although the *RMSD* indicates that the loading patterns of the models based on $\sigma^2 = 0.01$ are a bit more similar across subsamples (.026 for loadings, .035 for inter-correlations) than the models based on $\sigma^2 = 0.05$, these models are still rather similar across subsamples (.044 for loadings, .042 for inter-correlations, see Table 2). As Model 5 and 6 have a superior fit, further results are reported for these models.

The loadings and factor inter-correlations for Model 5 and Model 6 are given in Table 3. Although there are some differences between the loadings of the two models, the loading differences did not alter the meaning of the factors (i.e., the expected salient loadings remained on the expected factor for both sub-samples). Overall, the cross-loadings of variables representing gc on gf and BS were larger than the cross-loadings of variables representing gf. Although the models allowing for substantial cross-loadings have a superior fit, the size of the cross-loadings is not extreme. Only the cross-loading of *text comprehension (German)* on the BS factor was greater .30. This indicates that the cross-loadings do not modify the general meaning of the factors. The substantial factor inter-correlations indicate that a general factor representing the common variance of all factors is conceivable. The factor inter-correlations of Model 5 and 6 were similar. In both subsamples, the inter-correlations of the mean plausible values were considerably larger than the inter-correlations of the factors. Especially the correlation of the mean plausible values for M with the mean plausible values for gf and BS was extremely large. Previous research indicates that mean plausible values are similar to regression factor scores and it has been shown in simulation-studies that mean plausible values as well as regression factor scores may result in an over-estimation of the factor inter-correlations (Beauducel & Hilger, 2022a). In line with these previous findings, inter-correlations of mean plausible values were larger than the factor inter-correlations in the present study (see Table 3). In contrast, the inter-correlations of correlation-preserving mean plausible values correspond exactly to the factor inter-correlations of the BCFA factors presented in Table 3. Therefore, correlation-preserving mean plausible values result in unbiased loadings in second-order factor analysis.

Table 3. Model 5/Model 6 (subsample 1/2, prior $\sigma^2 = 0.05$), completely standardized loadings and factor inter-correlations

	gc	gf	M	BS
Grammar (German)	**.23 / .28**	.21 / .17	-	.22 / .24
orthography (German)	**.40 / .34**	.10 / .12	-	.04 / .07
text comprehens. (Ger)	**.16 / .21**	.18 / .16	-	**.34 / .33**
vocabulary (German)	**.21 / .29**	.24 / .20	-	.17 / .19
computer knowledge[a]	**.20 / .21**	.29 / .19	-	.14 / .22
grammar (English)	**.94 / .94**	-.09 /-.07	-	-.15 /-.19
communication (English)	**.69 / .76**	-.11 /-.09	-	.07 / .01
vocabulary (English)	**.75 / .78**	.05 / .00	-	-.11 /-.10
figural analogies	.06 / .08	**.45 / .37**	-	.10 / .16
verbal analogies	.09 / .10	**.48 / .41**	-	.09 / .16
arithmetic	.00 / .03	**.76 / .69**	-	.00 / .06
numerical estimation	.02 /-.02	**.71 / .72**	-	-.15 /-.14
relate information	.06 / .13	**.39 / .39**	-	.16 / .14
figural memory	-	-	**.52 / .49**	-.08 /-.07
verbal memory	-	-	**.50 / .51**	.18 / .20
compare numbers	-.02 / .00	-.01 /-.04	-	**.22 / .24**
social skills – customer	-.06 /-.02	-.16 /-.27	-	**.56 / .60**
organizational skills	-.06 /-.04	.06 / .06	-	**.43 / .47**
social skills – team	-.03 /-.02	-.12 /-.14	-	**.60 / .58**
social skills – error	-.04 /-.01	-.22 /-.28	-	**.58 / .57**
concentration	.01 / .00	.04 / .05	-	**.38 / .37**
factor inter-correlations				
gf	.49 / .50			
M	.54 / .59	.69 / .61		
BS	.59 / .62	.74 / .76	.74 / .68	
inter-correlations of mean plausible values				
gf	.72 / .70			
M	.76 / .76	.87 / .79		
BS	.77 / .78	.88 / .91	.91 / .84	

Note. Parameters of Model 5 are before the slash, parameters of Model 6 are behind the slash; gc = crystallized intelligence, gf = fluid intelligence, BS = basic skills/clerical speed, M = memory; cross-loadings of tasks on M were fixed to zero; expected salient loadings and cross-loadings of an absolute size greater than .30 are marked in boldface.

The g-factor loadings resulting from BCFA of mean plausible values and BCFA of correlation-preserving mean plausible values are given in Table 4, along with and the corresponding model fit. Whereas a one-factor model fits perfectly for the mean plausible values, the fit is less pronounced for the correlation-preserving mean plausible values. Thus, the over-estimated factor-intercorrelations of the mean plausible values (see Table 3) might suggest that there is not much more than g-variance in the data. However, the *RMSD* between the inter-correlations reproduced from the g-loadings and the original factor inter-correlations (the original factor inter-correlations are presented in Table 3) show that the mean plausible values result in g-loadings that do not represent the original factor inter-correlations (as indicated by high *RMSD* in Table 4). In contrast, the g-loadings based on correlation-preserving mean plausible values represent the original factor inter-correlations rather well (as indicated by low *RMSD* in Table 4).

Table 4. Factor-loadings of g based on BCFA of mean plausible values versus correlation-preserving mean plausible values

	mean plausible values		correlation-preserving mean plausible values	
	Subsample 1	Subsample 2	Subsample 1	Subsample 2
gc	.80	.80	.64	.68
gf	.92	.92	.82	.82
M	.95	.86	.83	.76
BS	.96	.98	.90	.92
ppp[a]	.054	.000	.000	.000
BIC	16586.56	17386.69	21634.01	21885.22
RMSEA	.046	.204	.069	.141
CFI	.999	.979	.996	.979
RMSD[b]	.190	.169	.016	.040

Note. [a]χ^2-based posterior predictor *p*-value, *BIC* = Bayesian Information Criterion; *RMSEA* = root mean square error of approximation; *CFI* = comparative fit index; [b]*RMSD* = root mean square difference between the inter-correlations reproduced by the g-loadings and the inter-correlations of the first-order factors.

Discussion

The main results of the present study can be summarized as follows: BCFA allows for the identification of factors representing fluid intelligence, crystallized intelligence, memory, and basic skills/clerical speed in a complex test battery that has been developed for the assessment of cognitive abilities in applied settings. The factor basic skills/clerical speed comprises speeded performance on moderately complex routine tasks and is therefore different from speed factors that are typically based on very simple speed tasks. This factor combines aspects of Carroll's (1993) broad speediness factor with broad processing speed. It represents several aspects of the cognitive speed domain (Wilhelm & Kyllonen, 2021) and might therefore be of special interest in the prediction of job performance (Sackett, Zedeck, & Fogli, 1988). In order to investigate the effect of non-zero cross-loadings, models without cross-loadings, models based on small prior variance of cross-loadings, and models based on moderate prior variance of cross-loadings were investigated. The models without cross-loadings had an inacceptable fit, although the factor inter-correlations were freely estimated and although two correlated errors were specified. Models based on small prior variance of cross-loadings had an improved fit, and models based on larger prior variance of cross-loadings had an even superior fit. This indicates that models without cross-loadings, sometimes termed 'independent clusters models', yield suboptimal representations of the factor structure of the test battery. This corroborates findings of simulation studies that BCFA may allow to overcome limitations of the independent clusters model (Xiao, et al., 2019). The patterns of expected salient loadings and cross-loadings as well as the factor inter-correlations were similar across a random split of the total sample into large subsamples. This indicates that the test battery allows for a robust measurement of intelligence factors that have been established in several studies (e.g., Carroll, 1993).

Overall, the factor-intercorrelations were rather large, which indicates that the test battery measures a relevant amount of general intelligence. Subsequent analyses based on mean plausible values indicate that the g-factor captures nearly all the measured variance. However, the analysis also shows that this results from an overestimation of the factor inter-correlations when the analysis is based on mean plausible values. In contrast, analyses based on correlation-preserving mean plausible values indicate that the amount of g-variance is substantial, but g does not represent the total common variance. Accordingly, we recommend to use correlation-preserving mean plausible values instead of mean plausible values for hierarchical factor analysis. In any case, the substantial amount of g-variance indicates that the test battery might have similar criterion validity as the tests that have been included in the meta-analyses by Salgado et al. (2003), Schmidt and Hunter (1998) and Sackett et al. (2021).

Moreover, it could also be that a systematic separation of content variance for verbal, numerical and figural intelligence, as it has been proposed by Guttman and Levy (1991), Guilford (1988), and Jäger (1984), might allow for an improvement of the discriminant validity of the factors. In order to establish a content facet for verbal,

numerical, and figural intelligence, additional marker variables (e.g., for numerical memory) should be added to the test battery. The possibility of a faceted model might be investigated in future studies.

As noted above, the true factor inter-correlations were overestimated by the mean plausible values. The overestimation of the inter-correlations was most pronounced for correlations with the memory factor. As this factor was only represented by two variables, it could be that the weak representation of the factor caused this effect. Because of the overestimation of the factor inter-correlations, mean plausible values cannot be recommended as an alternative to conventional scale scores for this test battery. However, the present results indicate that correlation-preserving mean plausible values may allow for an improvement of discriminant validity of scores representing the BCFA factors. As outlined in the introduction, computing correlation-preserving mean plausible values is a method to derive score estimates of the latent variables or factors resulting in path coefficients corresponding exactly to the path coefficients estimated by means of BCFA (Beauducel & Hilger, 2022b). Hence, they may be seen as especially useful when applying the results of BCFA to the factorial scoring of the test battery at hand. Each factor of the BCFA model can be represented by corresponding correlation-preserving mean plausible values (e.g., crystallized intelligence or fluid intelligence). So, as an important application of the present findings, the presented method can be used to derive estimates for a factorial scoring of a test battery, especially if the test was developed to meet practical demands so that cross-loadings are likely to occur.

Acknowledgement

This study was funded by the German Research Foundation (DFG), BE 2443/18-1.

References

Asparouhov, T., & Muthén, B. (2010a). Plausible values for latent variables using Mplus [Technical Report]. https://www.statmodel.com/download/Plausible.pdf

Asparouhov, T., & Muthén, B. (2010b). Bayesian analysis using Mplus: Technical implementation (Technical appendix). Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/BSEMFINAL10212011.pdf

Asparouhov, T., & Muthén, B. (2017). Prior-Posterior Predictive P-values. Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/PPPP.pdf

Asparouhov, T., & Muthén, B. (2022). Multiple Imputation with Mplus. https://www.statmodel.com/download/Imputations7.pdf

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al.. *Journal of Management, 41,* 1561–1577. https://doi.org/10.1177/0149206315591075

Beauducel, A. & Hilger, N. (2022a). Coefficients of factor score determinacy for mean plausible values of Bayesian factor analysis. *Educational and Psychological Measurement, 82*, 1069–1086. https://doi.org/10.1177/00131644221078960

Beauducel, A. & Hilger, N. (2022b). Correlation-Preserving Mean Plausible Values as a Basis for Prediction in the Context of Bayesian Structural Equation Modeling. *International Journal of Statistics and Probability, 11*, 1-11. https://doi.org/10.5539/ijsp.v11n6p1

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30,* 429–452. http://dx.doi.org/10.1207/s15327906mbr3003_6

DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen.* Berlin: Beuth.

Dombrowski, S.C., McGill, R.J., & Canivez, G.L. (2018). Hierarchical exploratory factor analyses of the Woodcock-Johnson IV Full Test Battery: Implications for CHC application in School Psychology. School Psychology Quarterly, 33, 235-250.https://doi.org/10.1037/spq0000221

Frey, M. C. (2019). What we know, are still getting wrong, and have yet to learn about the relationships among the SAT, intelligence and achievement. *Journal of Intelligence, 7*, 26. https://doi.org/10.3390/jintelligence7040026

Frey, M.C., & Detterman, D.K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science, 15*, 373–378. https://doi.org/10.1177/10.1111/j.0956-7976.2004.00687.x

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48(1)*, 1-4. https://doi.org/10.1177/001316448804800102

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence, 15(1)*, 79-103. https://doi.org/10.1016/0160-2896(91)90023-7

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253–270. http://dx.doi.org/10.1037/h0023816

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

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Structural research on intelligence: Competing models, new developments, perspectives]. *Psychologische Rundschau, 35(1)*, 21–35.

Jöreskog, K. G. & Sörbom, D. (2018). *LISREL 10 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100*, 399-416. https://doi.org/10.1037/a0037674

Lang, J. W. B., Kersting, M. & Beauducel, A. (2016). Hierarchies of factor solutions in the intelligence domain: Applying methodology from personality psychology to gain insights into the nature of intelligence. Learning and Individual Differences, 47, 37–50. http://dx.doi.org/10.1016/j.lindif.2015.12.003

Lang, J. W. B., Kersting, M., Hülsheger, U. R. & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology, 63,* 595-640. https://doi.org/10.1111/j.1744-6570.2010.01182.x

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111(3),* 490–504. https://doi.org/10.1037/0033-2909.111.3.490

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika, 46*, 337-341. https://doi.org/10.1007/BF02293740

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. https://doi.org/10.1016/j.intell.2008.08.004

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313-335. https://doi.org/10.1037/a0026802

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: Comments on, variations of, and extensions to Muthén and Asparouhov (2012). *Psychological Methods, 17*, 336–339. https://doi.org/10.1037/a0027130

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73(3), 482–486. https://doi.org/10.1037/0021-9010.73.3.482.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 07(11),* 20402068. https://doi.org/10.1037/apl0000994.

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13*, 150–170. https://doi.org/10.1037/1082-989X.13.2.150

Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*, 573-605. https://doi.org/10.1111/j.1744-6570.2003.tb00751.x

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274. https://doi.org/10.1037/0033-2909.124.2.262

Schmitz, F. & Wilhelm, O. (2019). Mene Mene Tekel Upharsin: Clerical speed and elementary cognitive speed are different by virtue of test mode only. *Journal of Intelligence, 7(16)*, 1-19. https://doi.org/10.3390/jintelligence7030016

Spearman, C. (1904). General intelligence, objectively determined and measured. The American Journal of Psychology, 15, 201-292. http://www.jstor.org/stable/1412107

Süß, H.-M., & Beauducel, A. (2015). Modeling the construct validity of the Berlin Intelligence Structure Model. *Estudos de Psychologia/Pychological Studies, 32(1),* 13-25. https://doi.org/10.1590/0103-166X2015000100002

Weide, A.C., Scheuble, V., & Beauducel, A. (2021). Bayesian and maximum-likelihood modeling and higher-level scores of interpersonal problems with circumplex structure. *Frontiers in Psychology. Quantitative Psychology and Measurement,* 12:761378.

Wilhelm, O. & Kyllonen, P. (2021). To predict the future, consider the past: Revisiting Carroll (1993) as a guide to the future of intelligence research. *Intelligence, 89*, 101585. https://doi.org/10.1016/j.intell.2021.101585

Xiao, Y., Liu, H., & Hau, K.-T. (2019). A Comparison of CFA, ESEM, and BSEM in Test Structure Analysis. Structural Equation Modeling: A Multidisciplinary Journal, 26, 665-677. https://doi.org/10.1080/10705511.2018.1562928Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. Structural Equation Modeling, 26(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

Appendix

Table A1. Models and parameter specifications

Model	factor variances	factor inter-correlations	expected salient loading	cross-loadings[a]
1	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	fixed (piors: μ = 0.00, σ^2 = 0.00)
2	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	fixed (piors: μ = 0.00, σ^2 = 0.00)
3	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	free (piors: μ = 0.00, σ^2 = 0.01)
4	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	free (piors: μ = 0.00, σ^2 = 0.01)
5	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	free (piors: μ = 0.00, σ^2 = 0.05)
6	fixed (piors: μ = 1.00, σ^2 = 0.00)	free[b]	free[b]	free (piors: μ = 0.00, σ^2 = 0.05)

Note. [a]No variability of cross-loadings on the factor M was allowed, the indications refer to the factors gc, gf, and BS; [b]free parameters have a prior mean of zero and a very large prior variance.

Example for Mplus syntax

```
Title: Model 5: Bayesian CFA for subsample1 with prior variance of 0.05;
  DATA: File is
  Multicheck_Wirtschaft_und_Administration_Mplus_labels_subsample1_scales.csv;

  LISTWISE=Off;

  VARIABLE:
  NAMES ARE
  Pbn
  age
  GramGE
  OrthGE
  TextGE
  VocGE
  ComKno
  ComDenk
  GraEN
  ComEN
  VocEN
  FRGram
  FRKomm
  FRWort
```

```
SoSk
Zahlver1
Compnum
FigAna
VerbAna
Calc
NumEst
FigMem
VerbMem
OrgSk
SoSkTe
SoSkEr
RelInf
Concen
;
        missing = all (999999);

USEVARIABLES
GramGE
OrthGE
TextGE
VocGE
ComKno
GraEN
ComEN
VocEN
SoSk
Compnum
FigAna
VerbAna
Calc
NumEst
FigMem
VerbMem
OrgSk
SoSkTe
SoSkEr
RelInf
Concen  ;

ANALYSIS:
ESTIMATOR = BAYES;
BCONVERGENCE = 0.005;
BITERATIONS = 400000;

MODEL:

gc by GramGE* OrthGE TextGE VocGE ComKno GraEN ComEN VocEN ;
gc@1;

gf by FigAna* VerbAna Calc NumEst RelInf  GraEN Compnum;
gf@1;

gc by Compnum* (gcCompn);
gc by FigAna*  (gcFigAn);
gc by VerbAna* (gcVerbAn);
```

```
gc by Calc*    (gcCalc);
gc by NumEst*  (gcNumEst);
gc by RelInf*  (gcRelInf);

gf by GramGE*  (gfGramGE);
gf by OrthGE*  (gfOrthGE);
gf by TextGE*  (gfTextGE);
gf by VocGE*   (gfVocGE);
gf by ComKno*  (gfComKno);
gf by ComEN*   (gfComEN);
gf by VocEN*   (gfVocEN);

M by FigMem* (1);
M by VerbMem* (1);
M@1;

BS by SoSk* OrgSk SoSkTe SoSkEr Concen ;
BS@1;


gf with BS (gfBS);
gc with BS (gcBS);
M  with BS (MBS);


gc by SoSk*   (gcSoSk);
gc by OrgSk*  (gcOrgSk);
gc by SoSkTe* (gcSoSkTe);
gc by SoSkEr* (gcSoSkEr);
gc by Concen* (gcConcen);

gf by SoSk*   (gfSoSk);
gf by OrgSk*  (gfOrgSk);
gf by SoSkTe* (gfSoSkTe);
gf by SoSkEr* (gfSoSkEr);
gf by Concen* (gfConcen);

BS by FigAna* (BSFigAn);
BS by VerbAna*(BSVerbAn);
BS by Calc*   (BSCalc);
BS by NumEst* (BSNumEst);
BS by RelInf* (BSRelInf);

BS by GramGE* (BSGramGE);
BS by OrthGE* (BSOrthGE);
BS by TextGE* (BSTextGE);
BS by VocGE*  (BSVocGE);
BS by ComKno* (BSComKno);
BS by GraEN*  (BSGraEN);
BS by ComEN*  (BSComEN);
BS by VocEN*  (BSVocEN);
BS by Compnum*(BSCompn);

BS by FigMem* (BSFigMem);
BS by VerbMem*(BSVerMem);
```

```
! Two correlated errors:
VerbMem with TextGE (VMTeGE);
NumEst with Calc (NumCal);


MODEL PRIORS:
gcCompn~N(0,0.05);
gcFigAn~N(0,0.05);
gcVerbAn~N(0,0.05);
gcCalc~N(0,0.05);
gcNumEst~N(0,0.05);
gcRelInf~N(0,0.05);

gfGramGE~N(0,0.05);
gfOrthGE~N(0,0.05);
gfTextGE~N(0,0.05);
gfVocGE~N(0,0.05);
gfComKno~N(0,0.05);
gfComEN~N(0,0.05);
gfVocEN~N(0,0.05);

gcSoSk~N(0,0.05);
gcOrgSk~N(0,0.05);
gcSoSkTe~N(0,0.05);
gcSoSkEr~N(0,0.05);
gcConcen~N(0,0.05);

gfSoSk~N(0,0.05);
gfOrgSk~N(0,0.05);
gfSoSkTe~N(0,0.05);
gfSoSkEr~N(0,0.05);
gfConcen~N(0,0.05);

BSFigAn~N(0,0.05);
BSVerbAn~N(0,0.05);
BSCalc~N(0,0.05);
BSNumEst~N(0,0.05);
BSRelInf~N(0,0.05);

BSGramGE~N(0,0.05);
BSOrthGE~N(0,0.05);
BSTextGE~N(0,0.05);
BSVocGE~N(0,0.05);
BSComKno~N(0,0.05);
BSGraEN~N(0,0.05);
BSComEN~N(0,0.05);
BSVocEN~N(0,0.05);
BSCompn~N(0,0.05);

BSFigMem~N(0,0.05);
BSVerMem~N(0,0.05);

gfBS~N(0,0.05);
gcBS~N(0,0.05);
MBS~N(0,0.05);
```

```
VMTeGE~N(0,0.05);
NumCal~N(0,0.05);

OUTPUT:
STANDARDIZED (STDYX); Tech8;
SAVEDATA:
FILE IS Model5_Bayesian_CFA_subsample1_prior_variance_SD05_plausible_values.txt;
SAVE = FSCORES (500);
```

Supplement 1: Description of tasks of the test battery

Table S1. Descriptions with number of items and time available for tasks of the test battery

Task	Description	Items	Time/Minutes
Grammar (German)	In a series of cloze tests (German texts) the missing word or the missing word ending must be entered correctly, if necessary.	20	5
Orthography (German)	In German texts misspelled words must be marked and corrected as a free input.	15	5
Text comprehension (German)	Different questions are asked about a text and the correct statements must be marked.	17	8
Vocabulary (German)	Synonyms or antonyms of German words must be selected from a list.	14	4
Grammar (English)	In a series of cloze tests (English texts) the missing word or the missing word ending must be entered correctly, if necessary.	15	5
Vocabulary (English)	Synonyms or antonyms of English words must be selected from a list.	12	6
Communication (English)	In various English conversations, questions or answers fitting the context must be selected from different options.	20	5
Computer knowledge	Questions concerning the use of computers and digital aids must be answered by selecting answer options.	12	15
Figural analogies	Laws must be inferred from figurative material and applied to find the correct solution.	14	10
Verbal analogies	Laws must be inferred from verbal material and applied to find the correct solution.	13	7
Arithmetic	Various mathematical problems presented as text or in tabular form must be solved with the aid of a calculator and note material.	15	15
Numerical estimation	Solutions to arithmetic problems must be estimated by mathematical reasoning (without the aid of a calculator and note material).	17	7.5
Compare numbers	Two series of numbers must be compared for each item, and it must be indicated whether they are identical or not.	72	4
Figural memory (learning / retrieval)	A series of pictograms is shown (learning phase). After an interference phase, pictograms are shown again, and it must be stated whether they occurred in the learning phase or not.	20	1.5 / 2

Task	Description	Items	Time/Minutes
Verbal memory (learning / retrieval)	A text is presented (learning phase). After an interference phase, the correct answers to questions about this text must be given by selection of answer options.	17	3 / 4
Concentration	The coordinates of a target in a grid must be entered as correctly and quickly as possible.	120	3
Relate information	This group of tasks is about linking information from a text and a table and answering different questions with the help of this information.	25	10
Organizational skills	In this planning task, several scheduling requests must be coordinated, considering certain conditions.	23	10
Social skills – customer service	In this situational judgment tests, the task is to accommodate and process customer concerns in a professional manner.	10	No constraint
Social skills – team	This situational judgment test is about coordinating jobs in a team and responding to challenges in a goal-oriented manner.	12	No constraint
Social skills –error handling	This situational judgment test is about dealing appropriately with errors and criticism and solving problems.	10	No constraint

Note. Items = number of scored items; Time = minutes available during the test execution (time constraint for solving the scored items)

Supplement 2: Illustration of task types - extract

This supplement serves to illustrate the task types, the instructions and the user interface for the different tasks. Stimuli are taken from the original preparation items. Each task begins with a rather simple preparation item without time constraint. Participants have to give the correct answer to the preparation item or choose to have the correct answer presented to them in order to be able to continue to the start of the scored tasks in the subtest. Note that there is only one preparation item used for all three situational judgment test szenarios for the measurement of social skills, as they all contain the same instrcutions and the same user interface. Also, note that typically the scored tasks are not as easy to solve as the those in the preparation items. Table S2 gives an overview of the content of the figures.

Table S2. Content of figures in this supplement.

Task	Figure
Grammar (German)	Figure S1
Orthography (German)	Figure S2
Text comprehension (German)	Figure S3
Vocabulary (German)	Figure S4
Grammar (English)	Figure S5
Vocabulary (English)	Figure S6
Communication (English)	Figure S7
Computer knowledge	Figure S8
Figural analogies	Figure S9
Verbal analogies	Figure S10
Arithmetic	Figure S11
Numerical estimation	Figure S12
Compare numbers	Figure S13
Figural memory (learning phase)	Figure S14
Figural memory (retrieval phase)	Figure S15
Verbal memory (learning phase)	Figure S16
Verbal memory (retrieval phase)	Figure S17
Concentration	Figure S18
Relate information	Figure S19
Organizational skills	Figure S20
Social skills	Figure S21

Note. The complete materials (including Figures 2 to 21) are available on request from the corresponding author.

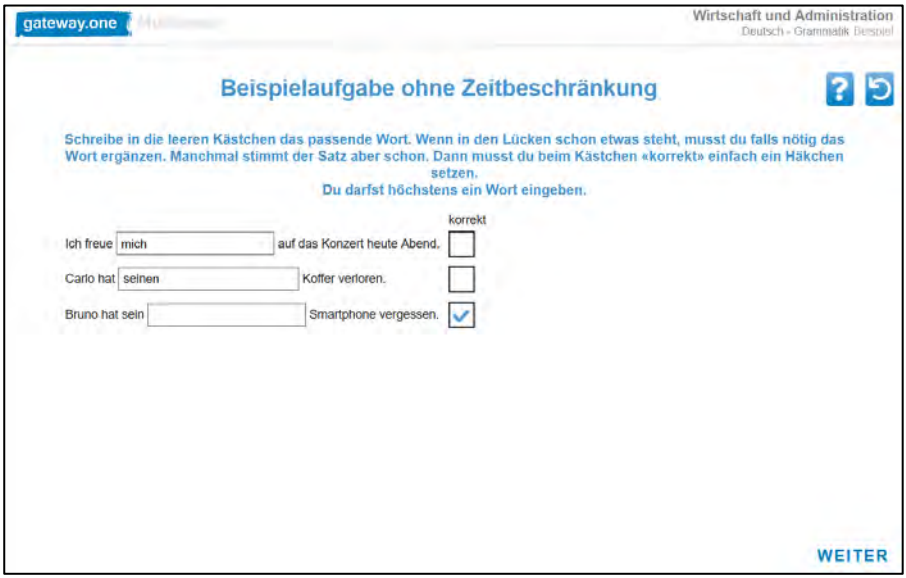

Figure S1. Grammar (German): simplified preparation item. Translation of instructions: “Write the appropriate word in the empty boxes. If something is already written in the gaps, you have to complement the word if necessary. Sometimes the sentence is already correct. In that case, just put a check mark in the "correct" box, and you can only write one word at the most.”