Original Article

Forced-Choice Versus Likert Responses on an Occupational Big Five Questionnaire

Luc Watrin[1], Mattis Geiger[2], Maik Spengler[1], and Oliver Wilhelm[2]

[1]HR Diagnostics, Stuttgart, Germany
[2]Institute of Psychology and Education, Ulm University, Germany

Abstract: Conventional self-report measures are prone to response biases, which distort measurement in any applied assessment. The forced-choice (FC) format was proposed as a potential remedy for these biases. The purpose of these studies was to develop and evaluate a FC questionnaire for the occupational context based on the five factor model of personality. A single-stimulus Likert questionnaire was contextualized for occupational settings and psychometrically optimized in Study 1 (N = 401). Considering optimal design strategies, we subsequently used this questionnaire to construct and validate a FC questionnaire in Study 2 (N = 517). Methodological add-ons to established approaches were applied to achieve decent confirmatory model fit. The new questionnaire shows good psychometric qualities and strong validity. We make suggestions for further applications and studies.

Keywords: forced-choice format, test development, Thurstonian IRT, five factor model, personality assessment, construct validity

Personality in the Occupational Context

The single-stimulus (SS) format, where respondents must rate statements individually on Likert-type scales, is the most widespread response format in personality tests. These tests are an integral part of many diagnostic processes, such as human resources (HR) processes (Benit & Soellner, 2013; König, Klehe, Berchtold, & Kleinmann, 2010; Piotrowski & Armstrong, 2006; Zibarras & Woods, 2010).

Despite their frequent usage, SS questionnaires have repeatedly been criticized for their susceptibility for different response distortions (e.g., socially desirable responding, response sets). A plethora of ways to deal with such distortions has been proposed (Alliger, Lilienfeld, & Mitchell, 1996; Donovan, Dwight, & Hurtz, 2003; Wetzel, Böhnke, & Brown, 2016; Ziegler, MacCann, & Roberts, 2011).

Yet, the reliance on SS personality tests is justified by results from several meta-analyses, which have shown that factors of the Big Five Factor model (FFM) of personality are valid predictors for a wide range of organizational outcomes (Barrick, Mount, & Judge 2001; Judge, Rodell, Klinger, Simon, & Crawford, 2013; Salgado, 1997; Schmidt, Oh, & Shaffer, 2016; Tett, Jackson, & Rothstein, 1991). A large body of research has shown additional ways of improving the predictive validity of generic personality tests even further, for example, by using contextualized items (e.g., Shaffer & Postlethwaite, 2012) and by using relevant lower-order facets of the Big Five (e.g., Dudley, Orvis, Lebiecki, & Cortina, 2006; Ones, Wiernik, Wilmot, & Kostal, 2016).

Forced-Choice

One of the most popular and promising alternative response formats for personality questionnaires is the forced-choice (FC) format, which has regained interest due to methodological developments in recent years (Brown & Maydeu-Olivares, 2011, 2012). In this alternative response format, multiple statements are grouped in blocks and respondents are asked to make decisions by comparing items within each block. For instance, in a block of three items, respondents may be instructed to rank order the items or choose which of the statements is "most like" and "least like" them. Using direct item comparisons instead of polytomous rating scales allows to reduce some problems of the latter, such as response distortions (e.g., Wetzel et al., 2016), reference group effects (e.g., Credé, Bashshur, & Niehorster, 2010), and non-linear distances between anchors (e.g., Wildt & Mazis, 1978; Worcester & Burns, 1975, as cited in Friedman & Amoo, 1999).

Regarding response distortions, extreme, midpoint, acquiescence, or disacquiescence response styles are unintentional distortions and reflect a respondent's systematic preference for a certain response category, generally independent of the item content (Wetzel et al., 2016). These

https://doi.org/10.1027/1614-0001/a000285

response styles vary between individuals. Ziegler (2015) cautions that test scores should not be compared among individuals when heterogeneity due to these distortions is left uncontrolled. The FC format inherently prevents such distortions, as it does not depend on a rating scale.

There is ample evidence that candidates can intentionally distort their scores in self-report measures to portray a different picture of themselves and that they will do so in high-stakes situations (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Viswesvaran & Ones, 1999). Widely used social desirability scales to control for faking have not proven to be an efficient method to approximate true test scores (Ellingson, Sackett, & Hough, 1999). Evidence for alternative approaches generally remained inconclusive (Ziegler et al., 2011).

The FC format allows the presentation of items of similar desirability together in one block. Therefore, they cannot be avoided successively like in the SS format, where it is possible to give each undesirable statement a low rating. This way, socially desirable responding could potentially be reduced (Guenole, Brown, & Cooper, 2018). Some studies have found evidence supporting this claim (Christiansen, Burns, & Montgomery, 2005; Martin, Bowen, & Hunt, 2002), yet some authors have even questioned the FC formats' ability to avoid social desirability on a general/theoretical level (e.g., see Meade, 2004).

While the discussed properties of the FC format regarding its resistance to unintentional and intentional distortions are highly favorable, the FC format has some drawbacks, too. First and foremost, the relative nature of FC scores has implications for their inter-individual interpretability and psychometric analysis. Conventional summative techniques from classical test theory (CTT) are not appropriate to analyze FC data, as they yield the same total score for every individual (Hicks, 1970; Meade, 2004). The resulting data is called *ipsative* (Cattell, 1944) and has several shortcomings, such as distorted construct and criterion validity, biased reliability estimates, and biased factor loadings in factor analysis (Brown & Maydeu-Olivares, 2013). Consequently, methods from CTT should never be used if inter-individual comparisons are to be made based on FC questionnaires.

Brown and Maydeu-Olivares (2011) have proposed an alternative statistical method, implementing Thurstone's *law of comparative judgment* (1927, 1931) in an item response theory (IRT) framework to circumvent the problems associated with conventionally scored FC data. The so-called Thurstonian IRT (T-IRT) model has successfully been applied to some FC tests so far, allowing the generation of normative scale scores and the achievement of substantial equivalence between FC and equivalent SS test versions, which is not obtainable when employing conventional methods (e.g., Anguiano-Carrasco, MacCann, Geiger, Seybert, & Roberts, 2015; Brown & Maydeu-Olivares, 2011, 2013; Joubert, Inceoglu, Bartram, Dowdeswell, & Lin, 2015). The results indicate that former limitations of the FC format (Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988) are no longer an issue. However, the number of studies evaluating the T-IRT model is still considered low.

Apart from methodological hurdles, the FC format can also lead to a higher cognitive load for test takers, as comparisons between items are more cognitively demanding than single ratings. Brown & Maydeu-Olivares (2011) therefore argue that blocks of 4 items might thus be the upper limit for practical use. Finally, especially in selection settings, the FC format might reduce the perceived influence of test takers' responses and could therefore reduce the acceptance toward the questionnaire as a fair diagnostic tool. However, since little empirical evidence exists concerning test takers' reactions toward FC, this issue should be considered more thoroughly in order to estimate the utility of the FC format.

Current Studies

The current research thus aims to extend the findings regarding the equivalence of FC and SS tests when using the T-IRT model, as well as its acceptance by test takers and therefore its utility in the applied context. In Study 1, we contextualized the items of an existing Big Five personality test with a facet structure to fit the occupational context, validated it in a first sample, and selected a psychometrically optimized short form. In Study 2, a FC test was developed using the items from the short form. The reliability, construct validities, and criterion-related validity of this test were evaluated in a second sample. A multi-trait-multi-method approach (MTMM) was used to relate the FC test to its SS equivalent, another Big Five personality test, as well as cognitive tests.

Study 1

Methods

Measure

A German version of the Big Five Aspects Scales (BFAS; DeYoung, Quilty, & Peterson, 2007; translated by Mussel & Paelecke, 2018) served as a basis for the following test developments. The scale consists of 100 items from the AB5C-IPIP Pool (Goldberg, 1999), which measure 10 distinct but correlated aspects of the FFM. A generic contextualization for the occupational setting was chosen to make the questionnaire widely applicable. As such, the original items were extended with references to the workplace

(e.g., "*At work* I am the first to act", "I inquire about *my colleague's* well-being") and the instructions were adapted to ask about typical behavior at work. The contextualization was conducted iteratively by several researchers and practitioners from the field of personnel selection and I/O psychology to reach an optimal solution. Finally, 50% of the items were adapted to achieve an adequate contextualization without repetition.

Sample

Participants were 401 young women and men who completed the test as part of a free online job-counseling program by the *Institut für Berufsprofiling* (HR Diagnostics AG). Participants were provided with individualized feedback regarding their results at the end of the test. As this job-counseling program is not related to any job offers or financial incentives at all, participants' first and foremost motivation and only benefit is to get honest feedback about their test performance. Therefore, mostly honest and motivated answers can be expected in this unproctored setting.

All items had to be answered on a 7-point Likert scale ranging from 1 = *very strongly disagree* to 7 = *very strongly agree*. Omitting items and revisiting previous items were not allowed. Six persons were removed due to zero variance in their test scores. The final sample consisted of 165 men and 230 women. The age of the sample ranged from 12 to 31 years ($M = 18.01$, $SD = 3.72$). With most participants being 15 years or older (88%), the sample largely reflects the age span of pupils applying for an apprenticeship, an integrated degree program, or an entry-level job in Germany.

Results

The descriptive statistics for the 10 aspects, as well as for the 5 factors, are shown in Table 1. McDonald's (1999) omega, which is an estimate for factor saturation and has been proven to be an adequate estimate for unidimensionality (Zinbarg, Revelle, Yovel, & Li, 2005; see also Revelle & Zinbarg, 2009), was above .70 for all scales and therefore acceptable. As scales were on the aspect level, we chose omega total over omega h. No serious deviations from the expected normal distributions were observed. The pattern of scale correlations was similar to those reported by DeYoung et al. (2007; table available from the corresponding author).

A confirmatory factor analysis (CFA) testing the structure with 10 primary and 5 secondary factors was performed in Mplus 7 (Muthén & Muthén, 1998–2010) using the MLR estimator. The model could not be computed due to problems involving several items from the politeness factor. Removing the factor altogether allowed for the computation of the model: $\chi^2(3{,}901) = 8{,}698$, $p < .001$, CFI = .610 and RMSEA = .056. Hence, goodness-of-fit criteria were poor according to established standards, specifically RMSEA $\leq$ .05 and CFI $\geq$ .95 (Hu & Bentler, 1999):

Table 1. Descriptive statistics, distribution parameters, and factor saturation of the BFAS full 100-item questionnaire in Study 1

	M	*SD*	Skewness	Kurtosis	ω_{total}
Neuroticism	3.73	1.54	0.10	0.70	
Volatility	3.59	1.49	0.02	0.46	.82
Withdrawal	3.88	1.59	0.10	0.53	.80
Agreeableness	5.03	1.47	−0.25	−0.13	
Compassion	5.06	1.48	−0.26	0.55	.85
Politeness	5.01	1.46	−0.23	0.16	.70
Conscientiousness	4.75	1.50	0.03	−0.43	
Industriousness	4.60	1.42	0.09	−0.23	.84
Orderliness	4.90	1.58	−0.13	−0.55	.83
Extraversion	4.75	1.46	−0.44	0.86	
Enthusiasm	4.88	1.44	−0.53	0.86	.80
Assertiveness	4.62	1.48	−0.20	−0.04	.85
Openness/Intellect	4.37	1.53	0.20	−0.20	
Intellect	4.45	1.45	0.01	0.18	.83
Openness	4.29	1.62	0.15	−0.57	.74

Using ant colony optimization (ACO) algorithms to design short scales of psychometric tests has proven to be a valid and efficient method to optimize psychometric properties, such as model fit and internal consistency (Olaru, Witthöft, & Wilhelm, 2015). ACO is a heuristic algorithm that identifies an optimal or close-to-optimal solution for combinatorial problems with solutions of differing quality over the course of iterations (Colorni, Dorigo, & Maniezzo, 1991). In the context of CFA, models with differing subsets of items are repeatedly tested until defined criteria are met. The latter can be specified freely, for example, minimizing RMSEA, maximizing CFI, maximizing the correlation with a criterion variable. The probabilities of items being selected for subsequent models are modified after each iteration, depending on how suitable the item was to reach the specified criteria. Over the course of iterations, the probability of items contributing to reach the specified criterion increases, leading to the selection of an optimal subset of items (Leite, Huang, & Marcoulides, 2008).

ACO is a new method in psychological research and the optimal specification of pheromone functions is a field of research on its own. To allow for comparisons and to keep as much consistency as possible with previous work, the optimization was specified with the criterion from Olaru et al. (2015), which maximizes the difference between CFI and RMSEA (Formula 1). We want to note that no gold standard for the optimization function has yet been established and that it must be adjusted or extended to the given situation depending on the research question, as is done in other recent applications (Olaru, Schroeders, Wilhelm, & Ostendorf, 2018; Schroeders, Wilhelm, & Olaru, 2016).

$$\text{Pheromone}_{new} = 0.9 \times \text{Pheromone}_{old} + (\text{CFI} - \text{RMSEA})_{best} \times 0.2 \times \text{IterationNum}. \quad (1)$$

As both the factor structure and the even distribution of items across factors had to be preserved and the final set of items had to be divisible by three to allow the construction of triplets in the FC questionnaire later, a 60-item solution was searched.

The R-script from Olaru et al. (2015) was adapted to perform ACO on the current model. The procedure implies repeated testing of alternate models in Mplus. Because the complete model containing all factors could not be computed, five separate optimization processes with one second-order and two first-order factors were performed. Each first-order factor was reduced to 6 items to obtain a final short scale with 60 items. As the algorithm uses a heuristic approach, it was applied 10 times to each factor in order to ensure that an optimal solution was found.

Table 2 shows that the model fit ranged from acceptable to good on the factor level. As all 10 runs per factor identified the same solution, the results can be deemed stable. ACO performs a purely numerical optimization, which is why the remaining items were checked to ensure content validity. The content of all scales was satisfying except for some items pertaining to the Politeness factor, which were quite harsh or extreme (e.g., "I take advantage of others."). While not problematic in a general context, those items seemed inappropriate for the occupational context and potentially difficult to integrate into a FC questionnaire. The items for this facet were therefore selected manually, to provide a scale that can be meaningfully applied in a selection context. Finally, the short scale with 60 items was modeled in a CFA, resulting in a model fit of $\chi^2(1,695) = 3{,}206$, $p < .001$, CFI = .715 and RMSEA = .048. Omega total for the shortened scales ranged from .72 to .81, except for Politeness ($\omega_{total} = .50$). We decided to keep the manually selected Politeness items to not alter the factor structure of the BFAS and to avoid imbalance in the triplet construction.

Table 2. Model fit and factor saturation of the shortened BFAS 60-item questionnaire. Models were estimated factor-wise

	E		C		N		O		A		Full model
CFI	.944		.957		.966		.937		.863		.715
RMSEA	.049		.049		.037		.049		.062		.048
	En	As	Ind	Or	Wi	Vo	Int	Op	Co	Po[a]	
ω_{total}	.75	.81	.80	.78	.76	.72	.80	.80	.80	.50	

Note. [a]Items of this aspect were chosen manually. E = Extraversion, C = Conscientiousness, N = Neuroticism, O = Openness (factor), A = Agreeableness; En = Enthusiasm, As = Assertiveness, Ind = Industriousness, Or = Orderliness, Wi = Withdrawal, Vo = Volatility, Int = Intellect, Op = Openness (aspect), Co = Compassion, Po = Politeness.

Conclusion

The full scale with translated and contextualized items showed adequate distributional properties, scale correlations, and internal consistency, which were comparable to those originally reported by DeYoung et al. (2007). A confirmatory factor model of the full item set could not be computed. Using ACO to select a reduced subset of items, acceptable to satisfying goodness-of-fit values could be attained on the factor level. For the full model, the RMSEA was good but other fit indices failed to reach conventional cut-off criteria for an acceptable model fit (Hu & Bentler, 1999). As Olaru et al. (2015) note, comprehensive self-report measures often suffer from poor model fit when tested with strict methods like CFA or IRT. While this reality is psychometrically unsatisfactory, inadequate fit values for broad FFM measures are generally the status quo (e.g., Borkenau & Ostendorf, 1990; Parker, Bagby, & Summerfeldt, 1993). As the fit criteria were good on the factor level and the internal consistency of the scales was generally acceptable, the selected short form was deemed appropriate for the following construction of a FC test.

Study 2

Methods

Stimulus Creation

The items from the short form questionnaire selected in Study 1 served as items for the development of the FC questionnaire. The 60 items were arranged into 20 triplets. As the number of items and facets did not allow for a complete balancing across triplets through permutation, an alternative design was developed using the AlgDesign package (Wheeler, 2014) in R version 3.2.3 (R Core Team, 2015). Thereby, a design was determined where each facet appears only once in each triplet, is compared at least once with every other facet, and the position of each facet within the triplets is balanced across the questionnaire. Additionally, the triplets were approximately balanced for desirability by considering the item means gathered in Study 1 as proxy to item desirability. For example, one triplet therefore consists of three items with the lowest mean value for their respective facet (see, e.g., Guenole et al., 2018, for an earlier application of this method). The keyed direction of items was not considered separately as they were expected to be reflected sufficiently in the item mean values. The final triplet design is depicted in Electronic Supplemental Material 1. Triplets were ordered as such in the study. Items are presented as codes, referring to DeYoung et al. (2007; Table 4); numbers are assigned facet-wise from top to bottom.

Figure 1. An example of a BFAS Forced-Choice item. Item order in a triplet could be changed using drag-and-drop with the mouse.

Sample

Study 2 was conducted on the same free online job-counseling platform used in Study 1. As the entire test battery was voluntary and quite lengthy (approx. 2 hr), many participants dropped out before reaching the end. The final sample consisted of 240 men and 277 women aged 12–31 years (M = 17.75, SD = 3.78).

Procedure and Measures

As in Study 1, all participants first had to complete a range of tests on the job-counseling platform. Among others, these included a short Big Five personality questionnaire for the work context (TAKE5; S & F Personalpsychologie, n.d.), a matrices test (MATRIX; S & F Personalpsychologie, n.d.), a test where arithmetic sequences had to be continued according to logic rules (KFM; S & F Personalpsychologie, n.d.), and a vocabulary test (LEXI; S & F Personalpsychologie, n.d.). At the end, both the BFAS Forced-Choice (BFAS FC) and the BFAS Single-Stimulus (BFAS SS) were presented. The BFAS SS had to be answered on a 7-point Likert scale. In the BFAS FC, items of a triplet were presented in boxes. Participants had to change the order of the items by putting the item best describing themselves on top of the list and the item describing themselves the least on the bottom. An example item is shown in Figure 1.

Each BFAS version was followed by a short acceptance questionnaire based on items from Kersting's (2005) *AKZEPT!-P*, a questionnaire assessing the acceptance of diagnostic measures.

Data Cleaning

Preceding the data analysis, stepwise data cleaning was performed. First, all participants with missing data were removed, as it indicates a premature dropout from the tests. Second, participants were excluded if they stated that they either did not understand the instructions or did not complete the BFAS FC or BFAS SS conscientiously. In the BFAS SS, additional participants were removed if their responses in a task had zero variance or were more than three standard deviations from the respective scale mean.

While a response to every item of the BFAS SS was mandatory, participants could move on to the next triplet of the BFAS FC at any time without altering the initial order, as the initial item order could coincidentally fit the participant's personal preferences.[1] To date, no guideline exists regarding how many altered triplets are realistic. Following a visual inspection of the amount of altered triplets across participants (Figure 2), a conservative cut-off value of 6 was selected, therefore excluding all participants with 5 or less (< 25%) altered triplets. After applying all cleaning steps, the remaining sample consisted of 517 subjects.

Results

Factor Structure of the 60-Item Short Form

First, as data-driven methods like ACO are vulnerable to overfitting and capitalization on chance, the short form selected in Study 1 was validated in the new sample. As such, a model with 5 secondary and 10 primary factors, as well as a model with 10 correlated primary factors, was tested. As Table 3 shows, the model fit was only marginally poorer for the hierarchical model (ΔCFI = −.015, ΔRMSEA = .005), which is positive evidence for the short form constructed in Study 1. The fit values for the correlated factors model were slightly higher and are relevant when comparing them to the FC model, as no approach for hierarchical models has been established there yet.

Factor Structure of the BFAS FC

The responses to the BFAS FC were first coded as binary outcomes, and the T-IRT model was fitted in Mplus after generating the necessary syntax with an Excel macro provided by Brown & Maydeu-Olivares (2012). The Mplus output of this model including a full list of parameters can be downloaded on Open Science Framework (https://osf.io/

[1] An optimized questionnaire design that requires participants to actively move the items of each triplet to a different part of the screen before ordering them avoids this problem and has successfully been applied in subsequent studies.

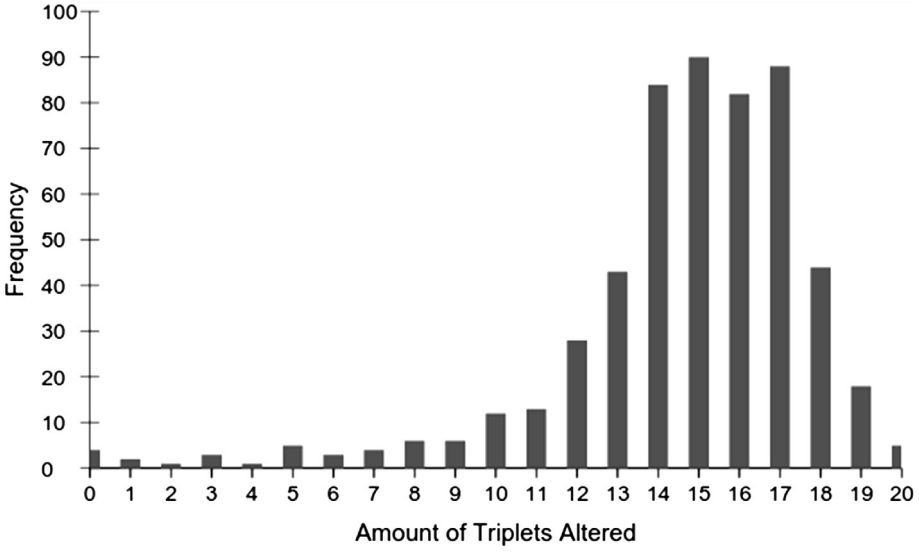

Figure 2. Amount of triplets altered by the participants (position of at least one item within a triplet altered).

Table 3. Fit values for the 60-item short form full model in Study 1 and Study 2

	df	χ^2	*p*	CFI	RMSEA
Study 1					
5 factors, 10 aspects	1,695	3,206	< .001	.715	.048
10 aspects	1,665	2,923	< .001	.763	.044
Study 2					
5 factors, 10 aspects	1,695	4,202	< .001	.700	.053
10 aspects	1,665	3,745	< .001	.751	.049

t8fgy/). Both distributions from the BFAS FC and SS overlapped substantially and showed only minor deviations from the normal distribution. The average of all trait scores ranged from θ = −0.86 to θ = 0.78 (*M* = 0.00, *SD* = 0.28) in the BFAS FC and from θ = −1.14 to θ = 1.63 (*M* = 0.00, *SD* = 0.34) in the BFAS SS.

As in other applications of the T-IRT model on FC data (Anguiano-Carrasco et al., 2015; Guenole et al., 2018), we observed empirical non-identification when estimating the model. In our case, the correlation matrix of the latent factors (ψ-matrix) was not positive definite and therefore no factor scores could be computed. One solution to this problem, which was applied in previous work (Anguiano-Carrasco et al., 2015; Guenole et al., 2018), is fixing values of the FC ψ-matrix to values derived from the SS ψ-matrix. However, because of method variance, these ψ-matrices can be expected to differ substantially from their true values and, therefore, the rationale of fixing covariances based on SS outcomes did not seem appropriate to us. An alternative to finding positive definite matrices is the Higham algorithm that searches for the closest positive definite matrix of a given matrix (Higham, 2002). We applied the Higham algorithm to our FC ψ-matrix in R version 3.2.3 (R Core Team, 2015) with the function nearPD of the Matrix package (Bates & Maechler, 2016), as well as the solution from previous work, namely fixing covariances to values derived from the SS ψ-matrix. In this iterative process, we first fixed all covariances and then reduced this number by one covariance per step via random selection, until a minimum number of mandatory fixations was identified. In our case, this resolved in two covariances fixed to values from the SS ψ-matrix. This iterative process is not optimized to finding an optimal solution, which the Higham algorithm is. Empirically, the Higham solution was superior to the iterative process, indicated by a better fit of the model (Higham solution: CFI = .878, RMSEA = .022; iterative process: CFI = .857, RMSEA = .024).

The Thurstonian IRT model was fit using the ULSMV estimator. The fit was as follows: $\chi^2(1{,}670) = 2{,}072$, $p < .001$, CFI = .878, RMSEA = .022. The mean intra-factor correlation of corresponding aspects was good (*M* = 0.530, *SD* = 0.247); however, the correlation of the aspects of extraversion was very low (*r* = .189).

Reliability

Reliability estimates for the BFAS SS were computed by applying a graded response model (Samejima, 1969). This model was fitted to the single-stimulus data using the mirt package (Chalmers, 2012) in R version 3.2.3 (R Core Team, 2015). Each factor with two underlying aspects was modeled separately.

While computing reliability estimates for single-stimulus data is an easy task, it is quite complex for FC data as the model is multidimensional and information of each item (i.e., result of a pairwise comparison) depends on two traits. Therefore, a test information curve cannot be computed as

Table 4. Reliability estimates for BFAS SS and BFAS FC

	SS	FC
Compassion	.908	.739
Politeness	.847	.678
Withdrawal	.924	.738
Volatility	.922	.753
Industriousness	.916	.739
Orderliness	.932	.718
Assertiveness	.902	.724
Enthusiasm	.901	.696
Intellect	.910	.788
Openness	.890	.745
M (*SD*)	.905 (.023)	.732 (.029)

Notes. Reliability estimates of SS and FC versions were estimated with different methods. SS = single-stimulus response format; FC = forced-choice response format.

Table 5. MTMM analysis with BFAS FC and BFAS SS on facet level

	M	*SD*	Min	Max
mThM	.62	.09	.42	.71
hTmM SS	.22	.16	.01	.62
hTmM FC	.35	.25	.02	.87
hThM	.21	.18	.00	.63
mFmM SS	.54	.30	.15	.97
mFmM FC	.53	.25	.19	.89
mFhM	.38	.16	.14	.65

Note. mThM = mono-trait-hetero-method, hTmM = hetero-trait-mono-method, hThM = hetero-trait-hetero-method, mFmM = mono-factor-mono-method, mFhM = mono-factor-hetero-method. Mono-factor correlations refer to correlations of aspects that belong to the same factor. SS = single-stimulus response format, FC = forced-choice response format. Min = minimum, Max = maximum.

normally. A practicable alternative is to use a simulation study approach, where true scores are simulated and correlated with simulated estimated scores to obtain an estimate of reliability (Brown & Maydeu-Olivares, 2011). Table 4 shows the reliability of both BFAS SS and FC as correlation between true and estimated scores.

The mean reliability of all aspects was higher for the BFAS SS (M = .905, SD = .023) than for the BFAS FC (M = .732, SD = .029). This finding was also the case for all corresponding scales. The difference between reliability estimates indicated that there was no systematic underestimation of a certain amount for the BFAS FC. All reliability estimates for the BFAS SS were above .80, those for the BFAS FC above .70, except for Politeness. The latter was already a problematic factor in Study 1 and had the lowest reliability in both tests.

Convergent Validity

To assess the convergent validity of the BFAS FC measure, we used trait scores in two subsequent multi-trait-multi-method (MTMM) analyses based roughly on Campbell and Fiske's (1959) work. Methods are reflected by SS and FC response formats, traits are the BFAS aspects. Traits in the first MTMM analysis are the aspects of the BFAS. First, means, standard deviations, minimum, and maximum of MTMM correlations were computed for the MTMM matrix of BFAS SS and BFAS FC. These indices are summarized in Table 5. In general, the following order of mean correlations is expected: reliabilities > validities (mono-trait-hetero-method [mThM]) > hetero-trait-mono-method (hTmM) > hetero-trait-hetero-method (hThM). Additionally, bearing the aspect and factor structure of the BFAS in mind, the intra-factor correlation within methods and between methods was examined in the first MTMM analysis. This analysis was expected to yield a mono-factor-mono-method (mFmM) greater than mono-factor-hetero-method (mFhM), which would be in turn greater than hThM.

Overall, the analysis revealed promising results. The mThM correlations are always expected to be the highest, which is the case in this analysis with a mean mThM correlation of $\bar{r}$ = .62. Additionally, its standard deviation was rather low (SD = .09) and the minimum value among the mThM correlations was still higher than the mean hTmM correlations for FC and SS. The mean hTmM correlation for both methods was low, with $\bar{r}$ = .35 for FC and with $\bar{r}$ = .22 even lower for SS. Here, greater indices of variance were found, but all of them were still acceptable. The hThM correlation is always expected to be lowest, which is the case with $\bar{r}$ = .21. In conclusion, this low value can be considered good.

The same pattern is found, when examining the correlations between aspects of factors. These relations within methods (mFmM) were expected to be similar and rather high, which is the case with $\bar{r}$ = .54 for SS and $\bar{r}$ = .53 for FC. Both have similar indices of scatter. The correlation within aspects of a factor but between methods (mFhM) is expected to be lower, than those within methods, but higher than hetero-trait correlations (hTmM and hThM). We found a mean mFhM correlation of $\bar{r}$ = .38, which fits the expected pattern exactly. The rather low standard deviation of SD = .16 supports the clarity of this pattern. Mean reliabilities were greater than all these values with $\bar{r}$ = .82.

We also tested these trends with Sawilowsky's (2002) I statistic. For this analysis, we added the reliabilities of the FC and SS versions as the reliability diagonal. When including mono-factor values to the test, we obtained I = 12 with p = .018. When excluding the mono-factor values, because they can be considered inadequate representatives for hTmM or hThM correlations, we obtained I = 10 and p = .008. Both tests are significant. Thus, the null hypothesis of Sawilowsky's I-test that all values are equal is discarded in favor of the alternative hypothesis: the values are in the correct order.

Table 6. MTMM analysis with BFAS FC, BFAS SS, and TAKE5, factor-wise only

	M	*SD*	Min	Max
mThM	0.51	0.14	.27	.70
hTmM SS	0.27	0.09	.15	.46
hTmM FC	0.38	0.23	.03	.69
hTmM T5	0.36	0.25	.06	.84
hThM	0.25	0.11	.02	.48

Note. mThM = mono-trait-hetero-method, hTmM = hetero-trait-mono-method, hThM = hetero-trait-hetero-method. SS = single-stimulus response format, FC = forced-choice response format. Min = minimum, Max = maximum.

Next, we conducted a second MTMM analysis by including the TAKE5 questionnaire in the MTMM matrix. Due to the factor structure of the TAKE5, which measures neither aspects nor facets, but only factors, we used factors as traits. Factor scores for BFAS SS and BFAS FC were computed by calculating the mean of the corresponding aspect scores. A summary of the second MTMM is presented in Table 6. Expectations regarding MTMM $\bar{r}$s were as in the previous analysis. As in the first analysis, the MTMM analysis resulted in expected relations.

The correlation within factors and between methods (mThM) revealed to be highest with a mean correlation of $\bar{r}$ = .51. Even the minimum r (r = .27) of mThM was higher than the average correlation between factors and methods (hThM; $\bar{r}$ = .25). The correlations between factors but within methods (hTmM) are expected to be between mThM and hThM, which was the case. For SS, hTmM was the lowest ($\bar{r}$ = .27). For FC ($\bar{r}$ = .38) and TAKE5 ($\bar{r}$ = .36), they were slightly larger, but all of them can be considered similar and appropriate in size.

Mean reliabilities were greater than all these values with $\bar{r}$ = .76. For the second MTMM analysis, we tested the trend with Sawilowsky's (2002) I statistic, too. We found I = 12 with p = .018. Again, this result indicates that the values are in correct order. Overall, the second MTMM analysis found the expected pattern, just as the first analysis did. Both full MTMM matrices are reported in Table A1 in Appendix A.

Divergent Validity

All performance tests were modeled with 2PL IRT models (Birnbaum, 1968) using the mirt package in R. The empirical reliabilities were r = .737 (matrices), r = .793 (numerical series), and r = .910 (vocabulary). The 2PL IRT models all fulfilled the criteria of CFI $\geq$ .95 and RMSEA $\leq$ .05 for a good model fit.

Correlations between BFAS SS, BFAS FC, and the performance tests are shown in Table 7. As expected, the correlations were generally low and near zero on average. The magnitude and direction of the correlations between both BFAS versions and the performance tests were generally comparable. The agreeableness factors showed surprisingly high correlations with the performance measures, and Openness showed the biggest difference between BFAS SS and FC.

Criterion Validity

To examine how test scores from both BFAS versions relate to external criteria, they were correlated with participants' self-reported average school grades. Self-reported school grades have several shortcomings, including restricted comparability and objectivity, but are nonetheless reliable (Sticca et al., 2017) and relevant variables in the educational and early occupational context and can therefore be considered a viable criterion in the young sample at hand. The German school grades were recoded, so that higher grades indicate better performance and subsequently z-standardized. Additionally, as Görlich and Schuler (2007) suggested, the z-scores were weighted to take the different levels of educational requirement of the German school

Table 7. Divergent validity of BFAS FC and SS with measures of intelligence

	Matrices		Numbers		Vocabulary	
	SS	FC	SS	FC	SS	FC
Compassion	.13	.20	.01	.12	.13	.24
Politeness	.23	.26	.14	.19	.18	.29
Withdrawal	−.06	−.04	−.07	−.08	−.01	−.04
Volatility	−.12	−.18	−.09	−.18	−.05	−.18
Industriousness	−.04	−.02	−.05	.03	−.07	−.02
Orderliness	−.09	−.09	−.15	−.11	−.15	−.15
Assertiveness	−.05	.02	−.01	.07	.01	.07
Enthusiasm	−.07	−.04	−.10	−.06	−.10	−.04
Intellect	.14	.15	.23	.15	.26	.17
Openness	.06	.24	−.09	.14	.12	.26
M (*SD*)	.01 (.11)	.05 (.15)	.02 (.11)	.03 (.12)	.03 (.13)	.06 (.16)

Note. All correlations > |.08| are significant on α = .05. SS = single-stimulus response format, FC = forced-choice response format.

Table 8. Criterion-related validity of SS and FC with school grades

	School grade depending on school type			
	SS	FC	Δr	$p_{\Delta r}$
Compassion	.18	.23	.05	.100
Politeness	.16	.21	.05	.140
Withdrawal	−.16	−.19	−.03	.200
Volatility	−.17	−.27	−.10	.003
Industriousness	.18	.21	.03	.180
Orderliness	−.01	.05	.06	.054
Assertiveness	.12	.21	.09	.007
Enthusiasm	.03	.05	.02	.310
Intellect	.34	.34	.00	.500
Openness	.15	.25	.10	.007

Notes. SS = single-stimulus response format, FC = forced-choice response format. All correlations in SS and FC > |.08| are significant at $\alpha = .05$. $p_{\Delta r}$ was estimated one tailed using Steiger's (1980) test for dependent correlations.

tiers into account ("Hauptschule" = −.60, "Realschule" = +.00, "Gymnasium" = +.50). As Table 8 shows, the correlations were highly similar and generally slightly higher for the FC version in particular. For Assertiveness, Volatility, and Openness, the difference in predictive validity was significant, indicating higher predictive validity of the FC version (significance was assessed using Steiger's (1980) test for dependent correlations).

Acceptance

Acceptance for the BFAS FC was generally lower than for its SS counterpart (Table B1 in Appendix B). The participants expressed that they felt restricted in their ability to answer adequately (M_{FC} = 3.94 vs. M_{SS} = 4.47, Appendix B) and gave the FC version a worse overall grade than the BFAS SS. All differences were statistically significant and of a small to medium effect size (Cohen, 1988). In absolute values, however, difference in the overall grade rating of the questionnaire was rather small: 2.38 (B- in the American school system) for the SS and 2.63 (C+, respectively) for the FC version. These overall grades are markedly better than those reported[2] for Raven's Progressive Matrices (M = 4.64; Bulheller & Häcker, 1998) by Kersting (2008) and similar to the global acceptance rating of M = 2.44 for the NEO-PI-R (Ostendorf & Angleitner, 2004) reported[2] by Steinmayr, Schütz, Hertel, and Schröder-Abé (2011).

Discussion

In two consecutive studies, we aimed to develop and evaluate a new forced-choice questionnaire assessing aspects (or facets) of the FFM in the occupational context. In Study 1, we therefore optimized an existing FFM questionnaire using ACO and reduced its item number from 100 to 60, while maintaining its factor and aspect structure. In Study 2, these items were grouped in triplets based on state-of-the-art methods to ensure balanced triplets regarding content and desirability. We evaluated the new questionnaires in terms of factor structure, reliability, convergent and divergent validity, criterion-related validity, and acceptance. For all criteria, the new questionnaire reached acceptable to good results for the given sample. Compared to the SS format, quality criteria of the FC format were mostly very similar or, for example, regarding model fit or predictive validity, even better. At the same time, response biases are limited in the FC format compared to the SS format. The forced-choice format can therefore be recommended for application. However, its analysis definitely requires an IRT approach, as it has shown to be the proper statistical way of dealing with FC computational issues described by Meade (2004).

Factor Structure

Self-report questionnaires often lack an acceptable model fit in confirmatory factor models, especially in the field of personality research. This problem is not new (e.g., Borkenau & Ostendorf, 1990; Parker et al., 1993), and recent approaches for dealing with this limitation have been described (e.g., Olaru et al., 2015). Even though we applied the same approach to the BFAS, namely ACO, the fit for the full model was still below the cut-off values suggested by Hu and Bentler (1999).

One possible explanation for the full model's fit is that we used a contextualized FFM questionnaire: one of the aspects, namely Politeness, contained relatively extreme items, which led to non-convergence of the full model. We could therefore not optimize the full model, but had to apply a factor-wise optimization, which might explain the lower fit of the full model. For future studies with BFAS items in the occupational context, one might rephrase the politeness items to less harsh content. On the other hand, model fit for individual factors was generally acceptable, and, as in previous FC work, the FC T-IRT model in this study revealed a substantially better fit in the full model than the SS version.

In FC modeling, the problem of non-positive definite matrices is usually dealt with by fixing covariances to values derived from the corresponding SS model. However, often the resulting FC ψ-matrix is similar but still different from the SS ψ-matrix and the latter therefore might not be the

[2] Reported acceptance ratings were recoded to fit the scale reported in this article (i.e., lower values equal higher acceptance and range from 1 to 6).

optimal solution to fixed values. Additionally, the SS ψ-matrix is derived from the SS model, which, as in our case, is of unsatisfying fit. Rather than relying on this prominent but somewhat limited approach, we applied the Higham algorithm as a methodologically sounder solution. Using the Higham algorithm to derive close covariances of the ψ-matrix that lead to a positive definite matrix successfully improved model fit and resulted in meaningful solutions. We therefore recommend this approach in order to reduce dependencies from a given SS model that is more often than not a suboptimal solution that FC aims to overcome.

Overall, the FC questionnaire had a fit that is rarely achieved in full FFM CFAs and was at least close to acceptable model fit by standard cut-off values. Fit indices for the SS model were considerably lower. However, as Guenole and colleagues (2018) already noted, even if the same items were used in the questionnaires, the models of FC and SS are not nested, because the indicators differ and a direct model comparison on a confirmatory level is thus not possible. Furthermore, the models differ in terms of the applied estimator (MLR vs. ULSMV or WLSMV) and relative amount of model restrictions. Yet, both questionnaires are expected to measure the same underlying psychological traits and alternative descriptive approaches, like comparing θ-distributions, yielded good results for the equivalence of both models.

Reliability and Validity

Compared to the SS version, the FC version had somewhat lower reliabilities. Yet, they were almost all acceptable as most reliabilities of aspects were > .7. While both estimates are operationalized as empirical reliabilities, it must be considered that the BFAS SS estimates were directly deduced from the data and that the BFAS FC estimates stem from simulations. While the lower reliability of the FC version can be considered a deficit, this phenomenon is not new. In defense of FC questionnaires, Brown and Maydeu-Olivares (2013) argue that, while a 7-point Likert scale contains six bits of information, a triplet only contains two bits of information per item, which might be the cause of reduced reliabilities. Overall, the FC reliability was acceptable.

We investigated construct validity via manifest MTMM analysis with rules of thumb based on descriptions by Campbell and Fiske (1959) and tests of trend (Sawilowsky, 2002) on factor and aspect levels. Descriptive MTMM statistics were in favor of construct validity and the tests of trend were significant. Overall, we can assume that the FC version has acceptable convergent validity. To assess divergent validity, we analyzed the questionnaire's relation to intelligence, which was generally low, just as in the SS version, and indicated good divergent reliability. The most considerable relation was that of the openness factors. This result is consistent with meta-analytic results that show substantial associations between cognitive abilities and openness (Ackerman & Heggestad, 1997). Several recent studies (e.g., Ackerman, 2009; Chamorro-Premuzic & Furnham, 2005; Zeidner & Matthews, 2000) also support our finding. Overall, we summarize that the newly developed FC questionnaire has decent construct validity.

As Anguiano-Carrasco et al. (2015), we also examined correlations with an external variable appropriate for our sample, namely school grades. Both BFAS versions showed high degrees of equivalence regarding direction and magnitude of these correlations. In three out of ten aspects, predictive validity was significantly higher for the BFAS FC. We conclude that there is some evidence for higher predictive validity in the FC version. However, further studies with additional criteria are necessary to replicate and extend these findings.

New to the field of modern FC questionnaires, yet of major importance for their application, is how test takers accept the questionnaire. While acceptance was lower for FC than for the SS version, in absolute values both versions were quite decently accepted on average and reached more acceptance than other most prominent psychological tests. This is a pleasant result, as the lengthy test battery that participants had to complete before reaching these questionnaires might even have reduced the overall mood and thereby the acceptance ratings of the forced-choice test, which was presented only after the assessment of cognitive ability and the TAKE5. In general, we interpret the acceptance to be high enough to apply the questionnaire. It is yet unclear whether the differences can be related to reduced possibilities of socially desirable responses. Consequently, different reactions are imaginable when applied in high-stakes settings, as the participants in this study completed the questionnaire voluntarily. Our work is meant to encourage further investigations on this research question.

Limitations and Implications for Future Research

The construction of the FC questionnaire was based on a relatively young sample, and the generalizability of the results is therefore limited to this population. The age range was relatively broad, but a range where personality traits are mostly stable (Borghuis et al., 2017; Olaru et al., 2018). Thus, the results are very promising and representative of generally young samples, on which a contextualized questionnaire would typically be applied. However, further validation is strongly recommended prior to application in other samples.

Great care was taken to create a balanced questionnaire design considering several key factors. Future studies may investigate whether social desirability can be balanced even better when extending the empirical approach described in this study. It might be beneficial to gather item desirability ratings explicitly (e.g., Konstabel, Aavik, & Allik, 2006) and consider them in the construction of the triplets, as Bäckström and Björklund (2013) showed that the desirability and the mean of an item are conceptually different.

Furthermore, the point that item means might not suffice as comprehensive information about an item's desirability is underlined by research that shows that candidates are able to identify criteria in assessments (ATIC; Klehe et al., 2012) and that consequently selective item adjustment is possible. Therefore, items with identical means but from different factors might still differ in their desirability in selection contexts. Thus, when striving to develop a test for a very different population, it is recommended to repeat all the steps in test development presented here for the different population. Yet, for the current situation and sample, item means can be considered a decent proxy to item desirability, as they reflect an item's attractivity (Guenole et al., 2018).

As discussed earlier, considering all these steps during the construction process might reduce the questionnaires' proneness to socially desirable responding or faking even further, especially in high-stakes situations. Yet, with every additional variable considered in the process of constructing triplets, balancing becomes more difficult and might at some point necessitate more items to allow for an optimal design (see Frey, Hartig, & Rupp, 2009, for an applied overview and Cochran & Cox, 1992, for a more general one).

Finally, misfit in the SS CFA model in Study 1 (i.e., correlated errors that were not estimated freely) could interfere with the triplet design. We found one case, where two items with a correlated error (modification index [MI] $\geq$ 10) ended up in one triplet. However, neither were this correlation when estimated freely in the SS CFA of Study 1 (r = .185, p = .001), nor any properties of these items in the forced-choice model (triplet distributions and loadings) obtrusive. Although such residual correlations were not an issue in the present studies, considering MIs and their distribution might be a valuable idea in future research.

On a more general level, the advantages of the FC format come at the expense of some challenges that must be kept in mind. First of all, IRT approaches like T-IRT require sample sizes that are considerably larger than those necessary for conventional CTT analyses. In situations of single person or small group assessments, using previously collected data and successively adding the data from new test takers might be a practicable interim solution until sufficient data from the sample at hand is gathered.

Summary

In two studies, we developed a FC questionnaire with adequate psychometric qualities to measure aspects of the FFM in an occupational context. Overall, while having some minor deficits that also apply to conventional personality questionnaires, our newly developed instrument performed well. The measure's reliability, validity, and acceptability in an occupational context ranged from acceptable to good. As the FFM is the most validated and dominant model of personality, and considering that the FC format is unsusceptible to response sets, we created a highly applicable questionnaire for young applicant samples. Applications in other samples, such as management-level applicants, require further validation.

Specifically new in this study, as compared to previous FC work, are firstly methodological add-ons, such as ACO for optimizing the underlying SS questionnaire, applying the Higham algorithm to achieve model convergence, and more external tasks for construct validation. Additionally, this questionnaire is the first forced-choice questionnaire contextualized for the occupational context directly based on the FFM of personality, including the aspect (or facet) level, and thereby capitalizes on previous recommendations for personality assessment. Considering all improvements to previous FC personality tasks and the high psychometric quality, we recommend this test for further application in research and applied assessment. For instance, future studies may investigate the acceptance of FC questionnaires in high-stakes samples, as well as their resistance against faking.

Availability

The developed questionnaire is available for research purposes via e-mail or ResearchGate from the corresponding author. Interested researchers might request three versions, namely the German original version as used in Study 2, a German revision with minor adjustments in the wording designed for broader selection processes, and an English translation of the questionnaire.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1614-0001/a000285

ESM 1. Table (.pdf)
The table shows the final triplet structure.

References

Ackerman, P. L. (2009). Personality and intelligence. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 162–174). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511596544.013

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219–245. https://doi.org/10.1037/0033-2909.121.2.219

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32–39. https://doi.org/10.1111/j.1467-9280.1996.tb00663.x

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment, 33*, 83–97. https://doi.org/10.1177/0734282914550387

Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology, 54*, 152–159. https://doi.org/10.1111/sjop.12015

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9–30. https://doi.org/10.1111/1468-2389.00160

Bates, D., & Maechler, M. (2016). *Matrix: Sparse and dense matrix classes and methods*. R package version 1.2-6 [R-package]. Retrieved from https://CRAN.R-project.org/package=Matrix

Benit, N., & Soellner, R. (2013). Scientist-practitioner gap in Deutschland: Eine empirische Studie am Beispiel psychologischer Testverfahren [Scientist-practitioner gap in Germany: An empirical study exemplified by psychological tests]. *Zeitschrift für Arbeits-und Organisationspsychologie A&O, 57*, 145–153. https://doi.org/10.1026/0932-4089/a000111

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Borghuis, J., Denissen, J. J., Oberski, D., Sijtsma, K., Meeus, W. H., Branje, S., … Bleidorn, W. (2017). Big Five personality stability, change, and codevelopment across adolescence and early adulthood. *Journal of Personality and Social Psychology, 113*, 641–657. https://doi.org/10.31234/osf.io/8pnvk

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11*, 515–524. https://doi.org/10.1016/0191-8869(90)90065-y

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*, 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*, 36–52. https://doi.org/10.1037/a0030641

Bulheller, S. & Häcker, H. (Eds.). (1998). *Raven's Progressive Matrices und Vocabulary Scales, von J. C. Raven, J. Raven und J. H. Court*. Frankfurt, Germany: Swets & Zeitlinger.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. https://doi.org/10.1037/h0046016

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*, 292–303. https://doi.org/10.1037/h0057299

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chamorro-Premuzic, T., & Furnham, A. (2005). *Personality and intellectual competence*. Mahwah, NJ: Erlbaum.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307. https://doi.org/10.1207/s15327043hup1803_4

Cochran, W. G., & Cox, G. M. (1992). *Experimental designs* (2nd ed.). New York, NY: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Colorni, A., Dorigo, M., & Maniezzo, V. (1991). Distributed optimization by ant colonies. *Proceedings of the First European Conference on Artificial Life, 142*, 134–142.

Credé, M., Bashshur, M., & Niehorster, S. (2010). Reference group effects in the measurement of personality and attitudes. *Journal of Personality Assessment, 92*, 390–399. https://doi.org/10.1080/00223891.2010.497393

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*, 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81–106. https://doi.org/10.1207/s15327043hup1601_4

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57. https://doi.org/10.1037/0021-9010.91.1.40

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155. https://doi.org/10.1037/0021-9010.84.2.155

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management, 9*, 114–123.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Görlich, Y., & Schuler, H. (2007). *Azubi-TH. Arbeitsprobe zur berufsbezogenen Intelligenz* [Azubi-TH. Work sample for measuring occupational intelligence]. Göttingen, Germany: Hogrefe.

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits:

Preliminary evidence from an application of Thurstonian item response modeling. *Assessment, 25*, 513–526. https://doi.org/10.1177/1073191116641181

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167. https://doi.org/10.1037/h0029780

Higham, N. J. (2002). Computing the nearest correlation matrix – a problem from finance. *IMA Journal of Numerical Analysis, 22*, 329–343. https://doi.org/10.1093/imanum/22.3.329

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

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162. https://doi.org/10.1111/j.2044-8325.1988.tb00279.x

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and Likert scale versions of a personality instrument. *International Journal of Selection and Assessment, 23*, 92–97. https://doi.org/10.1111/ijsa.12098

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875–925. https://doi.org/10.1037/a0033901

Kersting, M. (2005). *AKZEPT! Fragebogen zur Messung der Akzeptanz diagnostischer Verfahren* [AKZEPT! Questionnaire for measuring the acceptance of diagnostic procedures]. Aachen, Germany: Rheinisch-Westfälische Technische Hochschule Aachen.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests [On the acceptance of intelligence and performance tests]. *Report Psychologie, 33*, 420–433.

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*, 273–302. https://doi.org/10.1080/08959285.2012.703733

König, C. J., Klehe, U.-C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment, 18*, 17–27. https://doi.org/10.1111/j.1468-2389.2010.00485.x

Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality, 20*, 549–566. https://doi.org/10.1002/per.593

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an Ant Colony Optimization algorithm. *Multivariate Behavioral Research, 43*, 411–431. https://doi.org/10.1080/00273170802285743

Martin, B. A., Bowen, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247–256. https://doi.org/10.1016/s0191-8869(01)00021-6

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

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–551. https://doi.org/10.1348/0963179042596504

Mussel, P., & Paelecke, M. (2018). BFAS-G. Big Five Aspect Scales – German. In Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) (Ed.), *Elektronisches Testarchiv* (PSYNDEX Tests-Nr. 9007737). Trier, Germany: ZPID. http://doi.org/10.23668/psycharchives.2341

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A confirmatory examination of age-associated personality differences: Deriving age-related measurement invariant solutions using ant colony optimization. *Journal of Personality, 86*, 1037–1049. https://doi.org/10.1111/jopy.12373

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality, 59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

Ones, D. S., Wiernik, B. M., Wilmot, M. P., & Kostal, J. W. (2016). Conceptual and methodological complexity of narrow trait measures in personality-outcome research: Better knowledge by partitioning variance from multiple latent traits and measurement artifacts. *European Journal of Personality, 30*, 319–321. https://doi.org/10/bp27

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae. Testmanual* [NEO-PI-R: NEO personality inventory by Costa and McCrae. Test manual]. Göttingen, Germany: Hogrefe.

Parker, J. D., Bagby, R. M., & Summerfeldt, L. J. (1993). Confirmatory factor analysis of the Revised NEO Personality Inventory. *Personality and Individual Differences, 15*, 463–466. https://doi.org/10.1016/0191-8869(93)90074-d

Piotrowski, C., & Armstrong, T. (2006). Current recruitment and selection practices: A national survey of Fortune 1000 firms. *North American Journal of Psychology, 8*, 489–496.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154. https://doi.org/10.1007/s11336-008-9102-z

S & F Personalpsychologie. (n.d.). *KFM*, Unpublished test.

S & F Personalpsychologie. (n.d.). *LEXI*, Unpublished test.

S & F Personalpsychologie. (n.d.). *MATRIX*, Unpublished test.

S & F Personalpsychologie. (n.d.). *TAKE5*, Unpublished test.

Salgado, J. F. (1997). The Five Factor Model of personality and job performance in the European Community. *Journal of Applied Psychology, 82*, 30. https://doi.org/10.1037/0021-9010.82.1.30

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: Psychometric Society.

Sawilowsky, S. S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development, 35*, 78.

Schmidt, F. L., Oh, I. S., & Shaffer, J. A. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings* (Working paper). Retrieved from https://www.researchgate.net/publication/309203898

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One, 11*, e0167110. https://doi.org/10.1371/journal.pone.0167110

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*, 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Steinmayr, R., Schütz, A., Hertel, J., & Schröder-Abé, M. (2011). *Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz* [Mayer-Salovey-Caruso test of emotional intelligence]. Bern, Switzerland: Hans Huber.

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *PLoS One, 12*, e0187367. https://doi.org/10.1371/journal.pone.0187367

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology, 14*, 187–201. https://doi.org/10.1037/h0070025

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210. https://doi.org/10.1177/00131649921969802

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). New York, NY: Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0024

Wheeler, B. (2014). *AlgDesign: Algorithmic experimental design* [R-package]. Retrieved from https://CRAN.R-project.org/package=AlgDesign

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research, 15*, 261–267. https://doi.org/10.2307/3151256

Worcester, R. M., & Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *Journal of the Market Research Society, 17*, 181–197.

Zeidner, M., & Matthews, G. (2000). Intelligence and personality. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 581–610). New York, NY: Cambridge University Press. https://doi.org/10.1017/cbo9780511807947.027

Zibarras, L. D., & Woods, S. A. (2010). A survey of UK selection practices across different organization sizes and industry sectors. *Journal of Occupational and Organizational Psychology, 83*, 499–511. https://doi.org/10.1348/096317909x425203

Ziegler, M. (2015). "F*** you, I won't do what you told me!" – response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*, 153–158. https://doi.org/10.1027/1015-5759/a000292

Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.). (2011). *New perspectives on faking in personality assessment*. New York, NY: Oxford University Press.

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

History
Received October 10, 2017
Revision received August 27, 2018
Accepted August 27, 2018
Published online March 21, 2019

Acknowledgments
We thank Gabriel Olaru, Philipp Döbler, and Lianna Hrycyk for supporting the preparation of this article.

Authorship
Luc Watrin and Mattis Geiger contributed equally to this article and share the first authorship.

Mattis Geiger
Institute of Psychology and Education
Ulm University
Albert-Einstein-Allee 47
89081 Ulm
Germany
mattis.geiger@uni-ulm.de

Appendix A

Table A1. Multi-Trait-Multi-Method matrices of BFAS SS and FC on aspect level and BFAS SS and FC and TAKE5 on factor level

	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10	SS1	SS2	SS3	SS4	SS5	SS6	SS7	SS8	SS9	SS10
FC1	0.739																			
FC2	0.501	0.678																		
FC3	-0.356	0.218	0.738																	
FC4	-0.407	-0.080	0.886	0.753																
FC5	0.184	-0.035	-0.809	-0.676	0.739															
FC6	-0.231	-0.036	-0.062	0.055	0.549	0.718														
FC7	0.235	-0.200	-0.625	-0.572	0.485	-0.111	0.724													
FC8	0.606	-0.251	-0.533	-0.315	0.335	-0.020	0.189	0.696												
FC9	0.392	0.187	-0.733	-0.867	0.710	0.105	0.751	0.219	0.788											
FC10	0.635	0.365	-0.223	-0.438	0.032	-0.243	0.452	0.088	0.525	0.745										
SS1	0.575	0.297	-0.175	-0.171	0.139	0.002	0.100	0.347	0.178	0.301	0.908									
SS2	0.247	0.416	0.087	-0.002	-0.004	0.072	-0.188	-0.027	-0.006	0.098	0.452	0.847								
SS3	-0.212	0.038	0.668	0.651	-0.584	-0.097	-0.364	-0.292	-0.546	-0.103	-0.180	-0.082	0.924							
SS4	-0.221	-0.015	0.641	0.651	-0.572	-0.102	-0.333	-0.261	-0.545	-0.115	-0.214	-0.151	0.966	0.922						
SS5	0.109	-0.016	-0.529	-0.454	0.699	0.453	0.314	0.209	0.495	0.038	0.207	0.061	-0.622	-0.603	0.916					
SS6	-0.166	-0.125	-0.095	0.011	0.355	0.639	0.022	-0.009	0.070	-0.135	0.145	0.118	-0.166	-0.174	0.501	0.932				
SS7	0.182	-0.317	-0.528	-0.416	0.354	-0.064	0.646	0.334	0.442	0.215	0.213	-0.246	-0.336	-0.296	0.258	0.144	0.902			
SS8	0.314	-0.275	-0.442	-0.284	0.284	-0.002	0.281	0.589	0.218	0.094	0.422	0.056	-0.323	-0.313	0.225	0.175	0.642	0.901		
SS9	0.238	0.049	-0.544	-0.627	0.494	0.081	0.563	0.152	0.707	0.354	0.243	0.050	-0.451	-0.472	0.422	0.179	0.476	0.286	0.910	
SS10	0.301	0.198	0.050	-0.044	-0.074	-0.032	0.154	-0.020	0.144	0.564	0.249	0.058	0.040	0.030	0.009	0.020	0.038	0.009	0.148	0.905

mThM	mono aspect hetero method
hTmM SS	hetero aspect mono method SS
hTmM FC	hetero aspect mono method FC
mTmM	reliability
hThM	hetero aspect hereto method
mFmM	mono factor mono method
mFhM	mono factor hetero method

	A_SS	N_SS	C_SS	E_SS	O_SS	A_FC	N_FC	C_FC	E_FC	O_FC	take5_A	take5_N	take5_C	take5_E	take5_O
A_SS	0.878														
N_SS	-0.187	0.923													
C_SS	0.182	-0.456	0.924												
E_SS	0.151	-0.353	0.256	0.902											
O_SS	0.237	-0.290	0.244	0.300	0.903										
A_FC	0.523	-0.123	-0.065	-0.022	0.301	0.709									
N_FC	-0.082	0.678	-0.319	-0.477	-0.402	-0.190	0.746								
C_FC	0.072	-0.402	0.703	0.191	0.188	-0.031	-0.461	0.729							
E_FC	0.096	-0.409	0.201	0.663	0.369	0.158	-0.686	0.272	0.710						
O_FC	0.194	-0.384	0.160	0.309	0.671	0.523	-0.671	0.220	0.566	0.767					
take5_A	0.311	-0.482	0.261	0.030	0.211	0.274	-0.312	0.240	0.104	0.241	0.768				
take5_N	0.245	-0.588	0.345	0.187	0.182	0.166	-0.430	0.354	0.229	0.249	0.836	0.580			
take5_C	0.201	-0.321	0.481	0.300	0.443	0.101	-0.366	0.399	0.297	0.358	0.246	0.329	0.701		
take5_E	0.147	-0.283	0.263	0.618	0.252	0.018	-0.370	0.227	0.486	0.245	0.057	0.242	0.392	0.786	
take5_O	0.223	-0.215	0.309	0.302	0.508	0.177	-0.274	0.179	0.278	0.382	0.210	0.106	0.724	0.420	0.431

mThM	mono trait hetero method
hTmM SS	hetero trait mono method SS
hTmM FC	hetero trait mono method FC
hTmM T5	hetero trait mono method T5
hThM	hetero trait hetero method
mTmT	reliability

Note. SS = Single Stimulus response format; FC = Forced-Choice response format; FC1/SS1 = Compassion; FC2/SS2 = Politeness; FC3/SS3 = Withdrawal; FC4/SS4 = Volatility; FC5/SS5 = Industriousness; FC6/SS6 = Orderliness; FC7/SS7 = Assertiveness; FC8/SS8 = Enthusiasm; FC9/SS9 = Intellect; FC10/SS10 = Openness (aspect); A = Agreeableness; N = Neuroticism; C = Conscientiousness; E = Extraversion; O = Openness (factor).

Appendix B

Table B1. Comparison of acceptance ratings between BFAS FC and SS

	BFAS FC	BFAS SS	
	M (*SD*)	*M* (*SD*)	*d*
With the given response possibilities, I was able to express my actual attitudes to the statements mentioned.	3.94 (1.13)	4.47 (1.08)	−.48
Because of the given response options, I did not have the freedom to answer as it is accurate for myself.	3.68 (1.39)	3.15 (1.47)	.37
The given response options have forced me to make statements that do not correspond to my actual intentions.	3.59 (1.41)	2.62 (1.30)	.72
Which school grade would you give to the assessment just completed?	2.63 (0.89)	2.38 (0.97)	.27

Note. The original items used in the study were in German and are available from Kersting (2005). All differences are significant at $\alpha = .05$.

