Psychological Test and Assessment Modeling, Volume 58, 2016 (3), 439-454

The method of score estimation does not affect main results on gender differences in a Big Five short scale

Anja Schwibbe[1] *& André Beauducel*[2]

Abstract

We investigated the effect of score estimation method on the identification and quantification of gender differences in the Big Five. Group mean differences assessed with three different approaches were tested for significance (unit-weighted sum scores, latent variables in a MGCFA model, and latent variable score predictors). We administered the BF-16, a 16 adjective measure of the Big Five proposed by Herzberg and Brähler (2006), in a sample of 300 men and 273 women. We demonstrated gender-based measurement invariance by means of multi-group confirmatory factor analysis. Outcomes showed significant gender differences in Neuroticism, Openness to Experience, and Agreeableness, with women scoring higher in these dimensions regardless of scoring method. However, effect sizes for differences on latent variable level were more pronounced.

Keywords: Big Five, gender differences, latent variable, measurement invariance, multi-group confirmatory factor analysis

[1] *Correspondence concerning this article should be addressed to:* Anja Schwibbe, PhD, University Medical Center Hamburg-Eppendorf, Department of Biochemistry and Molecular Cell Biology, Martinistr 52, 20246 Hamburg, Germany; email: a.schwibbe@uke.de

[2] Institute of Psychology, University of Bonn, Germany

Introduction

Big Five

In the last twenty years five personality factors have been replicated in several cultures and languages (e.g. De Raad, Di Blas, & Perugini, 1998; Schmitt, Allik, McCrae, & Benet-Martinez, 2007), independently of inventories, statistical methods, and samples (Carroll, 2002; Fehr, 2006; Furnham & Fudge, 2008; Goldberg, 1990; John & Srivastava, 1999), although some other factors have also been found (e.g.; Ashton & Lee, 2007). The five basic dimensions Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C) allow for a description of several main differences between individuals (Caspi, Roberts, & Shiner, 2005; Fehr, 2006), whereby each factor summarizes a large number of distinct, more specific personality characteristics (Borghans, Duckworth, Heckman, & Bas ter, 2008). Neuroticism captures individual differences in emotional stability, and it can be defined with adjectives like anxious, irritable, and tense. Persons with high scores on the Extraversion-scale enjoy social interaction with others, are cheerful, and good entertainers. The scale Openness to Experience is associated with intellectual curiosity, love of variety, and lively fantasy life. Conscientiousness can be described in more detail with diligence, self-discipline, achievement striving, and dutifulness. Altruistic, in good faith and sociable are construct descriptive adjectives for Agreeableness (Costa Jr & McCrae, 1992).

Gender differences in the Big Five

There appear to exist systematic personality differences between various groups. A well-studied difference in psychological sciences is the personality difference between men and women. Most studies show that women appraise themselves to be more neurotic and agreeable (Costa Jr, Terracciano, & McCrae, 2001; Feingold, 1994; Lippa, 2010; Weller & Matiaske, 2009). These differences are very robust and can be captured with broad-band inventories like NEO-PI-R (Costa Jr et al., 2001), and Big Five Inventory (Lang, Lüdtke, & Asendorpf, 2001; Schmitt, Realo, Voracek, & Allik, 2008), as well as with several short instruments (e.g.; Goodwin & Gotlib, 2004; Rammstedt, 2007). The results for the traits C, E, and O are not as uniform. If O is measured with the NEO-PI-R on facet level, then the agreement depends on the reference object mentioned in the item. Males rather agree upon items concerning ideas and females upon items dealing with aesthetics, feelings, and actions. Concerning E, there is a clear dependence between gender and observed facets, too. Men tend to approve statements on Excitement Seeking, while women find themselves more in statements to Warmth. On dimension level, however, these facet scores are averaged and significant differences between males and females disappear (Costa Jr et al., 2001). In some studies women had slightly higher measures in C (Feingold, 1994; Goodwin & Gotlib, 2004; Rammstedt, 2007), but there are other studies finding no significant difference (Chapman, Duberstein, Sorensen, & Lyness, 2007; Costa Jr et al., 2001; Lippa, 2010; Weller & Matiaske, 2009). If studies

are conducted to draw conclusions for populations, an instrument must be able to represent these differences between various subgroups.

Testing measurement invariance in personality questionnaires

Before testing group differences, it is necessary to ensure that the instrument is measuring the same construct in the same way independently of group membership, which is called measurement invariance or measurement equivalence. Complete measurement equivalence or invariance indicates "…whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). This means that the same expressions of latent variables lead to the same values in the indicator variables, regardless of sample membership (Weiber & Mühlhaus, 2010, p. 232). The lack of this quality in turn is labeled with measurement bias or differential item functioning (DIF). Marsh (1994) recommends a demonstration of measurement invariance before group differences are reported as real, because DIF or lack of invariance can result in biased group comparisons (Desouky, Mora, & Howell, 2013).

Escorial and Navas (2007) describe several methods to determine DIF like standardization by calculating a standardized difference in proportions (Dorans & Holland, 1993), logistic regression calculations (Swaminathan & Rogers, 1990) or item-response theory based techniques such as Lords $\chi 2$ test. Under structural equation modeling (SEM), test measurement invariance or DIF can be investigated by means of multi-group confirmatory factor analyses (Jak, Oort, & Dolan, 2014; Marsh, 1994; Steenkamp & Baumgartner, 1998; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009; Vandenberg & Lance, 2000; Whitaker & McKinney, 2007). The several stages of measurement invariance are investigated with a cascade of comparisons of progressively restrictive models. Constraints of invariance (i.e., factor loadings, error variances, factor variances, and intercepts) are added stepwise and at each step the decrease of model fit is investigated (Vandenberg & Lance, 2000; Whitaker & McKinney, 2007). Using SEM may result in more accurate comparisons than comparison of unit-weighted sum scores (Woods, Oltmanns, & Turkheimer, 2009), because differences can be analyzed on level of latent variables, which are unaffected by measurement error.

Despite the recommendations and the variety of options of statistical analysis, there has been very few studies on gender-based measurement invariance of personality questionnaires (Memetovic, Ratner, & Richardson, 2014). Some studies investigated measurement invariance of scales measuring very specific constructs like risk for substance use (Memetovic et al., 2014), aggression (Ang, 2007), and perceived stress (Lavoie & Douglas, 2012). Marsh et al. (2010) criticized that, so far, for no Big Five inventory the a priori structure on item level could be proved with a Confirmatory Factor Analysis (CFA). However, in their investigations concerning the NEO-FFI they found gender-based measurement invariance. The following comparisons of latent means based on measurement equivalent models showed that women scored higher on all Big Five dimensions and especially on N and C. In the British Household Panel Study gender-based measurement invariance of a 15-item Five Factor Approach (FFA) instrument was ana-

lyzed (Marsh, Nagengast, & Morin, 2013). Based on invariant factor loadings and partially invariant item intercepts the authors investigated gender differences on the latent dimensions. Contrary to the results for the NEO-FFI (Marsh et al., 2010), women only had higher values in N, E and A, while men showed higher expressions in O. In C almost no gender difference was found (Marsh et al., 2013). Perhaps it is not possible to capture the full range of factor content with such a short instrument and thus, to detect gender differences.

Short versions of personality questionnaires

Against the background of convincing evidence of predictive validity of personality traits, more and more areas are interested in reliable and valid personality descriptions. For instance, application domains like consumer or political attitude research need personality measures only as an additional information source (Rammstedt & John, 2007). In large-scale surveys with many different questionnaires, or in longitudinal studies with several testing times, the application of short versions (Langford, 2003; Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001) could reduce boredom, frustration, fatigue, and drop-out rates (Burisch, 1984; Herzberg & Brähler, 2006). In the last 20 years, a variety of short versions to measure personality traits have been developed. Most of them use short statements, which have to be rated on Likert-scales. Briggs (1992), however, pointed out that sets of adjectives in a language are finite and definable, and that they can economically and easily be administered. Saucier and Goldberg (2001) also emphasized that "person-description and sedimentation of important differences in language both work primarily through the adjective function" (p. 850).

Looking for an appropriate short instrument to measure the Big Five in a validation study for student selection procedures, we became aware of the Big Five 16 adjective measure (BF-16) published by Herzberg and Brähler (2006). They started their test development with a German translation of the Ten Item Personality Inventory (TIPI; Gosling et. al, 2003). Yet, for several reasons, they did not recommend the use of this inventory as an alternative measure of the Big Five. Their results showed low internal consistencies, the scales were not normally distributed, and the convergences for O and A with the corresponding NEO-FFI scales were low. In order to develop an alternate, they decided to edit the original form, which consists of 10 items, whereby each item had two different describing adjectives. In the first data collection many participants criticized that sometimes it was not possible to assess two different adjectives with a single rating. To avoid this and to catch all information separately, they administered each adjective successively and added the label of the Big Five dimension itself as an adjective in four cases: conscientious, agreeable, neurotic, and introvert. The label for O "open to new experiences" was already part of the original form. Based on results of exploratory factor analysis, the new composed 24 adjective instrument was reduced to 16 items, because of too low primary and too high secondary loadings of some items. The subsequent CFA yielded reasonable fit indexes, conforming to the Big Five structure (Herzberg & Brähler, 2006). Reliability measures of internal consistency and test-retest-coefficients could also be improved by the extension, corrected item-total correlations were satisfactorily higher

than .30 and convergent and divergent correlations supported the validity of the instrument.

Aim of study

After measurement invariance of an instrument has been proved, gender differences can be investigated. There are, however, different possibilities for the investigation of gender differences. Steinmayr, Beauducel, and Spinath (2010) found that male and female differences in verbal, numerical, and figural intelligence depend on the method of score estimation. Most comparisons were based on unit-weighted sum scores, although latent variables allow for comparisons of the underlying pure content factors. Many decisions in Clinical Psychology, such as group assignments based on personality scores, rely on unit-weighted sum scores, although they are contaminated with measurement error. Latent variable models have the disadvantage that they do not provide individual scores, which are needed for individual diagnosis and assessment. Calculation of latent variable score predictors based on latent models can minimize this problem (Grice, 2001), but one has to bear in mind that because of indeterminacy latent variable score predictors and latent variables are not necessarily identical. On the other hand, it is impossible to calculate individual scores representing the latent variables themselves (Mulaik, 2010). Therefore, it is necessary to calculate latent variable score predictors whenever individual scores have to be compared. Depending on estimation method, it is possible to calculate different latent variable score predictors for one latent variable model (Steinmayr et al., 2010).

Against this background, it seems useful to compare male and female values by means of these three different methods: 1) unit-weighted sum scores, 2) latent variables and 3) latent variable score predictors. As precondition we investigate gender-based measurement invariance of BF-16 by means of MGCFA modeling.

Method

Participants

A total of 573 students dealt with a 16 adjective measure of Big Five (BF-16) as part of a validation study of student selection procedures. 300 of them were men with a mean age of 23.4 years (SD = 2.94; range: 19-38 years). The 273 participating women had also a mean age of 23.4 years (SD = 3.96; range: 18-44 years). With 398 subjects, the majority of our sample studied educational sciences or psychology (men: N = 160; women: N = 238), and the remaining 175 participants studied business or engineering sciences (men: N = 140; women: N = 35). All subjects completed the questionnaires voluntarily and gave written informed consent.

Instruments

The participants completed the BF-16 of Herzberg and Brähler (2006), which was developed to measure the Big Five by means of a short list of 16 adjectives, which are given in Table 1. The adjectives were rated on a 7-point Likert scale from 1 (disagree strongly) to 7 (agree strongly).

Analyses

We performed all analyses of the manifest variables and the latent variable score predictors with SPSS 22, and for structural equation modeling we used Mplus 6.1 (Muthén & Muthén, 1998-2010). First, we examined by means of confirmatory factor analysis (CFA) whether the proposed model by Herzberg and Brähler (2006) could be replicated in our sample. We also calculated a first-order model with correlated factors, and fixed the factor variances to unit variance. By virtue of statistically significant Kolmogorov-Smirnov-Z-Scores, all scales deviated from the normal distribution, and therefore we used the robust Maximum Likelihood Method of Satorra and Bentler (1994) as the test developer did. To assess the goodness-of-fit of the resulting models we reported the chi-square-statistic, the comparative fit index (*CFI*), the Tucker-Lewis-Index (*TLI*) and the root mean square error of approximation (*RMSEA*) and standardized root mean square residual (SRMR). Before analyzing gender differences with different scoring methods we used MGCFA to investigate gender-based measurement invariance. According to Vandenberg and Lance (2000) there are no strict guidelines for the order of making invariance constraints in the models. Our approach is based on the sequence of steps used by Gustavsson et al. (2008):

1. In a first step, we calculated a CFA model for both male and female participants separately.
2. The baseline CFA model to test configural invariance assumes that zero-loadings are found in the same location of the loading matrix in both groups (although the non-zero loadings need not to be equal across groups). All parameters as non-zero factor loadings, intercepts and residual variances in the two groups were freely estimated. Factor means were fixed to zero.
3. For investigating metric invariance factor loadings were constrained to equality for both groups, whereas intercepts and residual variances were freely estimated and factor means were fixed to zero. Change of model fit indexes was assessed to show that relationships of items and latent variables are the same.
4. Proving scalar invariance means to demonstrate that additionally constraining intercepts to equality did not deteriorate the model fit significantly. Factor means in the male group were fixed to zero and freely estimated in the female group.

To evaluate the change of model fit, we decided not to use the chi-square difference test (values are reported) because the chi-square test is very sensitive in large samples and very small differences become significant (Marsh, 1994). Therefore some authors argue

this method could be problematic and inappropriate (Brown, 2006; Marsh, 1994). We use fit indexes like Gustavsson et al. (2008): SRMR, RMSEA, CFI, and TLI. If the fit indexes of the more restricted model were on the same level like the indexes of the previous model measurement invariance is assumed.

However, as stated above, Steinmayr et al. (2010) showed that gender differences in intelligence were dependent on the method of score estimation. Therefore, we also expected an effect of the type of analysis on gender differences and tested for gender differences using three different approaches after establishing measurement invariance. First, we tested gender differences at the level of unit-weighted sum scores using t-tests. Secondly, we compared the latent variables in the gender-based measurement invariant MGCFA model and thirdly, we compared latent variable score predictors between men and women by means of t-tests. Since we expected correlated factors, we decided to calculate the latent variable score predictors by means of McDonald´s (1981) correlation preserving factor score estimates. To evaluate the impact of the different scoring methods, effect sizes for the observed gender differences are reported. We chose to report effect size *r*, calculated with *t*-values or Cohen's *d* (Rosenthal & Rosnow, 1991) for all comparisons.

Results

Descriptive statistics

The item and scale statistics of the BF-16 are reported in Table 1. Compared to the original our data showed lower coefficients alpha, particularly for the scales O and C, and all items measuring O had corrected item-total correlations lower than .25. The means and standard deviations were comparable with the results of Herzberg and Brähler (2006). We do not report gender-specific scale means at this point, since measurement invariance was not investigated until now. At first sight, the gender-specific reliability coefficients of N and O and corrected item-total correlations of N-items and O-items differed between men and women.

In a next step, we performed a CFA to investigate whether the model proposed by Herzberg and Brähler (2006) fits our data. In the first-order model the factors were allowed to correlate and the factor variances were fixed to unit variance. The fit was $\chi^2_{SB}(94) = 390.546$, $p < .001$, $CFI = .85$, $TLI = .81$, $SRMR = .07$ $RMSEA = .07$ with a 90% confidence interval of .067 - .082. The model is given in Figure 1. All loadings were above .30 with an average of .61. The factor inter-correlations ranged between -.41 to .45 with seven correlations reaching statistical significance (marked in Figure 1).

Table 1:
Psychometric Properties of the BF-16

	Total sample					Gender specific values			
Scale/Item-No.	*M*	*SD*	α	r_{it}	*K-S*	$\alpha_{(m)}$	$r_{it(m)}$	$\alpha_{(f)}$	$r_{it(f)}$
Neuroticism	3.10	1.09	.68		.09	.56		.70	
*1. easily upset**				.43			.28		.48
*2. anxious**				.42			.18		.51
3. calm^{+}				.50			.44		.51
4. emotionally stable^{+}				.53			.50		.43
Extraversion	4.51	1.53	.87		.09	.86		.88	
5. reserved^{+}				.72			.69		.74
6. quiet^{+}				.81			.70		.76
7. introverted^{+}				.73			.80		.82
Openness to Experience	5.23	.88	.32		.08	.26		.37	
8. open to new experiences				.18			.18		.18
9. complex				.22			.17		.27
10. uncreative^{+}				.17			.11		.23
Agreeableness	5.61	1.02	.67		.17	.62		.67	
11. sympathetic				.51			.45		.51
12. warm				.51			.45		.51
Conscientiousness	5.58	.85	.62		.12	.63		.62	
13. dependable				.42			.38		.48
14. self-disciplined				.47			.50		.45
15. disorganized^{+}				.35			.48		.45
*16. conscientious**				.47			.37		.34

Note. N_{total} = 573. K-S: Kolmogorov-Smirnov Z-value, all $p < .001$.
$_{m}$ = male specific values (N = 300).
$_{f}$ = female specific values (N = 273).
$^{+}$ denotes reverse-scored items.
*denotes removed items affected by DIF

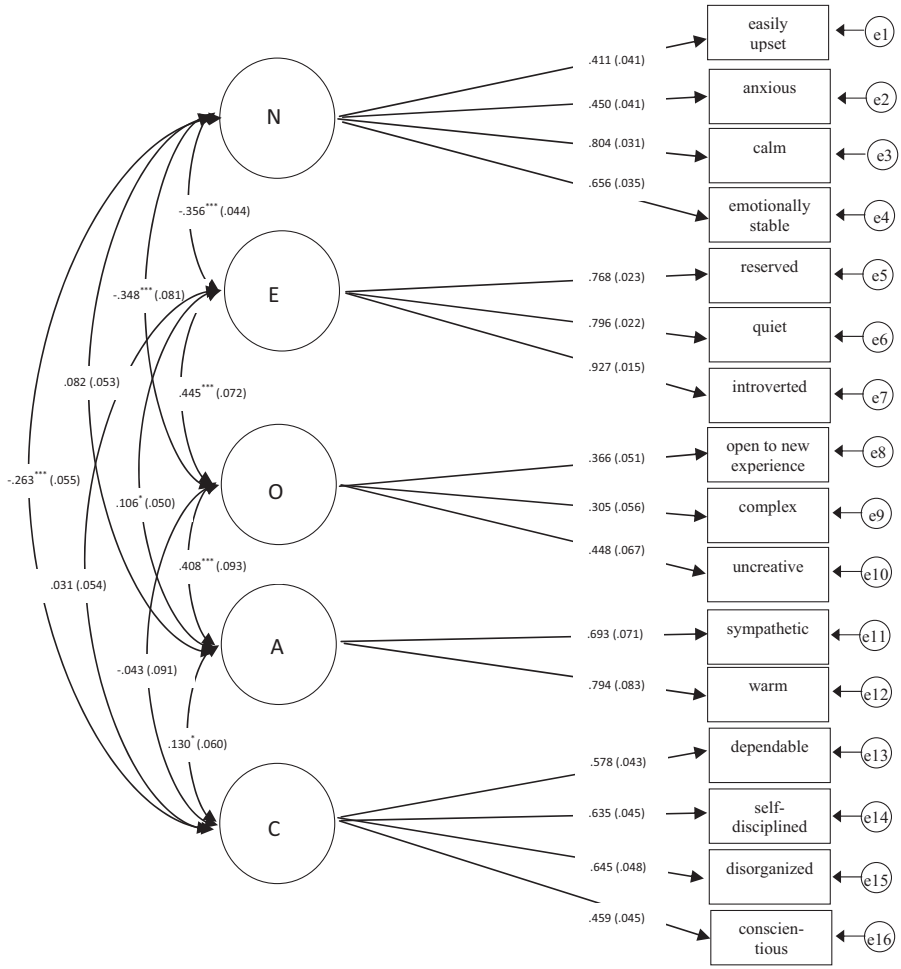

Figure 1:
CFA basic model of the BF-16. Significance of factor correlations is marked with $^{*}p \leq .10$, $^{***}p \leq .001$ (two-tailed.)

Testing for Differential Item Functioning

Testing for measurement invariance by means of MGCFA means to compare the fit indexes of different models with increasing equality constraints of parameters. Only SRMR values of all models met the usual cutoff recommendations for good fit (based on Hu & Bentler, 1999), the remaining fit indexes RMSEA, CFI, and TLI of all models did not. Comparing the different models with respect to measurement invariance we observed that the fit indexes of the different models summarized in Table 2 were not substantially changed by increasing the parameter constraints, confirming scalar invariance of the BF-16. Thus, comparison of group means can be executed for the latent variables (Temme & Hildebrandt, 2009).

Table 2:
Confirmatory factor analyses and tests of measurement invariance of the BF-16

	χ^2	df	SRMR	RMSEA	CFI	TLI
Confirmatory factor analyses in separate groups						
Male	234.01	94	.074	.070	.85	.81
Female	214.72	94	.067	.069	.88	.85
Test of Measurement Invariance						
1) Configural Invariance (equal form)	448.86	188	.070	.070	.87	.83
2) Metric Invariance (equal factor loadings)	503.56	204	.079	.072	.84	.82
3) Scalar Invariance (equal intercepts)	536.04	215	.082	.072	.83	.81

Note. χ^2 is the Satorra–Bentler χ^2. *df* = degrees of freedom. *SRMR* = standardized root mean square residual. *RMSEA* = root mean square error of approximation. *CFI* = comparative fit index. *TLI* = Tucker–Lewis index.

Gender differences and scoring method

All results regarding the comparisons of scoring method are summarized in Table 3. First, we performed a t-test with unit-weighted sum scores for each Big Five dimension. Women reached higher values in all scales, and in the dimensions N, O, and A the differences were statistically significant. Based on the formula of Rosnow et al. (2000) we calculated effect size *r* from *t*-values. According to Cohen's guidelines (1988) the differences in N and A should be regarded as weak effects. Actually, the difference in O was significant, but effect size r was too low to be considered as a weak effect.

In the gender-based measurement invariant MGCFA model latent variable means of male group were fixed to zero and means of female group were freely estimated. We found significant gender effects in the dimensions N, O, and A. In our calculations gender was encoded using “1” for male and “2” for female. The significant positive latent means indicated that women had significantly higher values in the relevant dimensions. To determine effect size *r* we followed Hancock (2001) who defined effect size *d* as a standardized difference between two means. Since one latent mean in the model was

fixed to zero and variances to one, the formula simplified so far that the estimated latent means corresponds to effect size *d*. Using the formula of Rosenthal (1991) we converted *d* into effect size *r*. Differences in N and A can be regarded as medium and in O as a weak effect (see Table 3).

Based on the loadings of the gender-based measurement invariant MGCFA model, latent variable score predictors were calculated by means of McDonald´s correlation preserving factor score predictors. Gender differences in the five latent variables (factors) were ascertained by means of *t*-tests. We also observed significant differences in N, O, and A between male and female participants. The inspection of effect size *r* for these differences yield a medium effect for N, a weak effect for O, and a medium effect for A.

Table 3:
Gender difference testing

	unit-weighted sum scores					MGCFA model			latent variable score predictors				
	M_m	SD_m	M_f	SD_f	r	est.	*S.E.*	r	M_m	SD_m	M_f	SD_f	r
N	2.77	1.3	3.58	1.2	.11***	.90	.10	.41***	-.316	.9	.347	1.0	.33***
E	4.50	1.5	4.52	1.5	.00	.01	.08	.01	-.002	1.0	.002	1.0	.00
O	5.13	.8	5.34	.9	.01*	.41	.15	.20**	-.107	1.0	.118	1.0	.11**
A	5.29	1.1	5.96	.8	.11***	.90	.12	.41***	-.330	1.0	.363	.8	.35***
C	5.68	.8	5.76	.8	.03	.08	.10	.04	-.021	1.0	.023	1.0	.02

Note. * $p \leq .10$, ** $p \leq .01$ *** $p \leq .001$ (two-tailed). $N_{total} = 573$. $N_{male} = 300$. $N_{female} = 273$. $M_{f/m}$ = mean in female/male group. $SD_{f/m}$ = standard deviation in female/male group. r = effect size *r*. est. = estimate of latent variable mean. *S.E.* = standard error.

Discussion

The influence of three different scoring methods on the identification and quantification of gender effects was investigated in the BF-16. Measurement invariance was proved by means of model comparisons with multi group CFA. In contrast to the results of Steinmayr et al. (2010) on intelligence, we found significant gender differences in N, O, and A independently of the scoring method. In unit-weighted sum scores, latent variables in a MGCFA model, as well as in McDonald´s correlation preserving factor score predictors, women had the higher values in N, O, and A. Effect sizes consistently revealed a medium effect for N and A, and a weak effect for O when mean-differences were determined for latent variables. The effect size for differences in unit-weighted sum scores yielded smaller values. It is therefore advisable to make group mean comparisons on level of latent variables to avoid influence of measurement error.

Our results concerning N and A were in line with the existing literature. Women appraise themselves to be more neurotic and agreeable than men (e.g.; Costa Jr et al., 2001; Lippa, 2010). Marsh et al. (2013) examined personality gender differences after proving

measurement invariance in a 15 item instrument. In accordance with their results of significantly higher values for women in N and A, we were also able to show that these differences occur when a short version of the Big Five is used and measurement invariance is established.

The research literature on O indicates a dependence of significant gender differences with the reference objects that are mentioned in the item (Costa Jr et al., 2001). Our results showed significantly higher values for women independently of the scoring method. Probably, it is not possible to consider different reference objects with items which only consist of a single adjective. It is therefore unlikely that differences in the reference object are the reason for gender differences.

To ensure that reported differences were real differences we tested our instrument for measurement equivalence prior to implementation of group comparisons. By means of MGCFA we could prove, that constraining factor loadings and intercepts to be equal for male and female participants did not deteriorate model fit indexes substantially. We concluded that scalar invariance can be assumed for the BF-16, which is required as prerequisite for group mean comparisons.

Fit indexes of our basic CFA model were all slightly below those of Herzberg and Brähler (2006), who following a statement from Raykov (1998) evaluated their model as reasonable. In particular the factor inter-correlations were more pronounced in our data. Marsh (1994) points out that the fit of the baseline model is very important, because otherwise the following comparison models could also not fit. However, he acknowledged that particularly in personality research this procedure could be too demanding and recommended to test for measurement invariance if the fit is not too bad.

Although we proved measurement-invariance for the BF-16, we discourage from using this short version as a proxy for a longer Big Five questionnaire especially due to bad item and scale statistics of the factor O. Corrected item-total correlations and coefficient alpha were so low that it is not justifiable to determine an individual trait value with these three items. The basic CFA model did not indicate sufficient construct validity of the BF-16 and this questionnaire should be revised with new or perhaps more items.

The present study has some limitations. Our data collection was carried out in a university setting with students from different fields. This results in a range restriction in age, and all participants have an above-average level of education. Furthermore, the distribution of male and female students in the various fields was very stereotypical. In humanities there were more women, while business or engineering sciences was mainly studied by men. It would be advisable to repeat the method comparison in a less restricted sample.

In summary, the lack of scoring method effects on gender differences in N, A, and O confirms previous research concerning gender differences in certain personality traits. Women seem to be more neurotic, agreeable and open to new experiences as compared to men. Furthermore, the missing method effect is an indication that these differences are so robust that they might cover real differences.

References

Ang, R. P. (2007). Factor structure of the 12-item aggression questionnaire: Further evidence from Asian adolescent samples. *Journal of Adolescence, 30*, 671-685.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 60*, 253-293.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Bas ter, W. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources, 43*, 972-1059.

Briggs, S. R. (1992). Assessing the Five-Factor Model of Personality Description. *Journal of Personality, 60*, 253-293.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford.

Burisch, M. (1984). You Don't Always Get What You Pay for: Measuring Depression with Short and Simple versus Long and Sophisticated Scales. *Journal of Research in Personality, 18*, 81-98.

Carroll, J. B. (2002). The five factor model of personality how complete and satisfactory is it. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 97-126). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; US.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: stability and change. *Annual Review of Psychology, 56*, 453-484.

Chapman, B. P., Duberstein, P. R., Sorensen, S., & Lyness, J. M. (2007). Gender Differences in Five Factor Model Personality Traits in an Elderly Cohort: Extension of Robust and Surprising Findings to an Older Generation. *Personality and Individual Differences, 43*, 1594-1603.

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

Costa Jr, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor inventory. Professional Manual.* Odessa, FL. : Psychological Assessment Resources.

Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322-331.

De Raad, B., Di Blas, L., & Perugini, M. (1998). Two independently constructed Italian trait taxonomies: Comparisons of among Italian and between Italian and Germanic Languages. *European Journal of Personality, 12*, 19-41.

Desouky, T. F., Mora, P. A., & Howell, E. A. (2013). Measurement invariance of the SF-12 across European-American, Latina, and African-American postpartum women. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 22*, 1135-1144.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Escorial, S., & Navas, M. J. (2007). Analysis of the Gender Variable in the Eysenck Personality Questionnaire Revised Scales Using Differential Item Functioning Techniques. *Educational and Psychological Measurement, 67*, 990-1001.

Fehr, T. (2006). Big Five. Die fünf grundlegenden Dimensionen der Persönlichkeit und ihre dreißig Facetten [Big Five. The five basic dimensions of personality and their 30 facets]. In W. Simon (Ed.), *Persönlichkeitsmodelle und Persönlichkeitstests: 15 Persönlichkeitsmodelle für Personalauswahl, Persönlichkeitsentwicklung, Training und Coaching* (pp. 113-135). Offenbach: GABAL Verlag GmbH.

Feingold, A. (1994). Gender Differences in Personality. A Meta Analysis. *Psychological Bulletin, 116*, 429-456.

Furnham, A., & Fudge, C. (2008). The Five Factor Model of Personality and Scales Performance. *Journal of Individual Differences, 29*, 11-16.

Goldberg, L. R. (1990). An Alternative Description of Personality. The Big Five Factor Structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goodwin, R. D., & Gotlib, I. H. (2004). Gender differences in depression: the role of personality factors. *Psychiatry research, 126*, 135-142.

Grice, J. W. (2001). Computing and Evaluating Factor Scores. *Psychology Methods, 6*, 430-450.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373-388.

Herzberg, P. Y., & Brähler, E. (2006). Assessing the Big-Five Personality Domains via Short Forms. *European Journal of Psychological Assessment, 22*, 139-148.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*, 117-144.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling, 21*, 31-39.

John, O. P., & Srivastava, S. (1999). The Big-Five Trait Taxonomy History, Measurement, and Theoretical Perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102-138). New York: The Guilford Press.

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen. [Validity and psychometric equivalence of the German version of the Big Five Inventory in young, middle-aged and old adults]. *Diagnostica, 47*, 111-121.

Langford, P. H. (2003). A one-minute measure of the Big Five? Evaluating and abridging Shafer's (1999a) Big Five markers. *Personality and Individual Differences, 35*, 1127-1140.

Lavoie, J., & Douglas, K. (2012). The Perceived Stress Scale: Evaluating Configural, Metric and Scalar Invariance across Mental Health Status and Gender. *Journal of Psychopathology and Behavioral Assessment, 34*, 48-57.

Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: testing evolutionary and social-environmental theories. *Archives of sexual behavior, 39*, 619-636.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling: A Multidisciplinary Journal, 1*, 5-34.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A New Look at the Big Five Factor Structure Through Exploratory Structural Equation Modeling. *Psychological Assessment, 22*, 471-491.

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement Invariance of Big-Five Factors Over the Life Span: ESEM Tests of Gender, Age, Plasticity, Maturity, and La Dolce Vita Effects. *Developmental Psychology, 49*, 1194-1218.

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika, 46*, 337-341.

Memetovic, J., Ratner, P. A., & Richardson, C. G. (2014). Gender-based measurement invariance of the Substance Use Risk Profile Scale. *Addictive Behaviors, 39*, 690-694.

Mulaik, S. A. (2010). *Foundations of factor analysis*. 2nd ed. New York: CRC Press.

Muthén, L. K., & Muthén, B. O. (1998-2010). *MPlus User´s Guide*. 6th ed. Los Angeles, CA: Muthén & Muthén.

Rammstedt, B. (2007). The 10-Item Big Five Inventory. *European Journal of Psychological Assessment, 23*, 193-201.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203-212.

Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences, 24*, 291-293.

Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality Correlates of Self-Esteem. *Journal of Research in Personality, 35*, 463-482.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. 2nd ed. New York: McGraw-Hill.

Rosnow, R. L., Rosenthal, R., & Rubin., D. B. (2000). Contrasts and Correlations in Effect-size Estimation. *Psychological Science, 11*, 446-453.

Satorra, A., & Bentler, B. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: SAGE.

Saucier, G., & Goldberg, L. R. (2001). Lexical Studies of Indigenous Personality Factors: Premises, Products and Prospects. *Journal of Personality, 69*, 847-879.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations. *Journal of Cross-Cultural Psychology, 38*, 173-212.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*, 168-182.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*, 78-107.

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence, 38*, 101-110.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity, 43*, 599-616.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*, 361-370.

Temme, D., & Hildebrandt, L. (2009). Gruppenvergleiche bei hypothetischen Konstrukten - Die Prüfung der Übereinstimmung von Messmodellen. [Group comparisons of hypothetical constructs - The verification of compliance of measurement models]. *Zeitschrift für betriebswirtschaftliche Forschung*, 138-185.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*, 4-70.

Weiber, R., & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS [Strucural Equation Modeling]*. Berlin: Springer Verlag.

Weller, I., & Matiaske, W. (2009). Persönlichkeit und Personalforschung. Vorstellung und Validierung einer Kurzskala zur Messung der „Big Five“. [Personality and Human Resource Management Research. Introduction of a Short “Big Five” Measurement Tool]. *Zeitschrift für Personalforschung, 23*, 258-266.

Whitaker, B. G., & McKinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods, 39*, 502-509.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*, 320-330.