Controlling acquiescence bias in measurement invariance tests

Aichholzer, Julian

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Aichholzer, J. (2015). Controlling acquiescence bias in measurement invariance tests. *Psihologija*, *48*(4), 409-429. https://doi.org/10.2298/PSI1504409A

PSIHOLOGIJA, 2015, Vol. 48(4), 409–429

UDC 303.64
159.9.072
DOI: 10.2298/PSI1504409A

Controlling acquiescence bias in measurement invariance tests

Julian Aichholzer

Department of Methods in the Social Sciences, University of Vienna, Austria

Assessing measurement invariance (MI) is an important cornerstone in establishing equivalence of instruments and comparability of constructs. However, a common concern is that respondent differences in acquiescence response style (ARS) behavior could entail a lack of MI for the measured constructs. This study investigates if and how ARS impacts MI and the level of MI achieved. Data from two representative samples and two popular short Big Five personality scales were analyzed to study hypothesized ARS differences among educational groups. Multiple-group factor analysis and the random intercept method for controlling ARS are used to investigate MI with and without controlling for ARS. Results suggest that, contrary to expectations, controlling for ARS had little impact on conclusions regarding the level of MI of the instruments. Thus, the results suggest that testing MI is not an appropriate means for detecting ARS differences per se. Implications and further research areas are discussed.

Keywords: measurement invariance; acquiescence; multiple-group factor analysis; Big Five

Quantitative social and behavioral research frequently relies on the technique of self-report instruments, a collection of questionnaire items that aim to measure the respondents' attitudes or personality, i.e., latent constructs. An important cornerstone in assessing the items' psychometric validity is, inter alia, the practical equivalence of construct measurements, known under the heading of *measurement invariance* (MI)[1] (Meredith, 1993): respondents equally interpret the question/request with regard to the construct and equally make use of the response scale (e.g. Chen, 2008). Note that achieving certain levels of MI of the instrument is a vital prerequisite for meaningful comparisons of correlations between construct scores and their mean scores across respondents,

Corresponding author: julian.aichholzer@univie.ac.at

Acknowledgements: This research is based on the author's doctoral dissertation conducted under the auspices of the Austrian National Election Study (AUTNES), a National Research Network (NFN) sponsored by the Austrian Science Fund (FWF) [S10902-G11]. A previous version was presented at the European Conference on Psychological Assessment (ECPA), Zurich, July 2015. I would also like to thank the anonymous referees for helpful comments.

1 Note that MI should not be confused with the term *modification indices*.

since these parameters are otherwise erratic and potentially biased (see Chen, 2008; Guenole & Brown, 2014; Steinmetz, 2013).

Technically speaking, MI means that observed scores in indicators (or items) measure the same latent constructs (or factors) and equally relate to those constructs in different contexts (i.e., across respondent groups). This hypothesis can, for instance, be tested statistically by means of multiple-group factor analysis (MG-FA) (Jöreskog, 1971; though see also Kankaraš & Moors, 2010).

The impact of acquiescence response bias on measurement invariance

In this study I address a key problem in assessing MI, namely that self-report instruments "notoriously" suffer from *systematic* measurement bias in observed scores (Podsakoff, MacKenzie, & Podsakoff, 2012). In particular, a crucial source of bias in popular rating scales is the acquiescence or agreeing response style (hereafter ARS) to statements including stimuli about approval or agreement (Bentler, Jackson, & Messick, 1971; Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Podsakoff et al., 2012; Rammstedt & Farmer, 2013). ARS also means that respondents consistently tend to endorse both a regular (or pro-trait) and a negatively phrased (or con-trait) item (Paulhus, 1991), while this behavior could be careless responding (Krosnick, 1991) or mere acceptance of inconsistent self-descriptive attributes (Bentler et al., 1971). ARS is, nevertheless, considered to be a behavior largely consistent across domains and stable over time (Billiet & Davidov, 2008; Danner, Aichholzer, & Rammstedt, 2015; Weijters, Geuens, & Schillewaert, 2010; Wetzel, Lüdtke, Zettler, & Böhnke, 2015), thus allowing potential control for this tendency when analyzing the data.

The problem with ARS or systematic bias in scale usage is that it violates the assumption of MI, which is also defined as *unbiasedness* of the indicator-construct relationship (e.g. Millsap & Meredith, 1992). Measurement bias is thus said to occur if respondents exhibit variation in response outcomes that is not only due to the level of the hypothesized traits to be measured (e.g. personality), but also due to a violating factor such as ARS. According to this conjecture, ARS interferes with measurement validity and can bias measurement parameters that are the basis for conducting statistical MI tests (i.e., factor loadings or item intercepts in MG-FA).

Previous research

Previous research has identified several conjectures with regard to the biasing impact of ARS in terms of violating MI. First, ARS is known to entail spurious correlations between questionnaire items and, hence, the true item-factor loading structure becomes more blurred with increasing levels of ARS (Aichholzer, 2014; McCrae, Herbst, & Costa, 2001; Podsakoff et al., 2012; Rammstedt & Farmer, 2013; Rammstedt, Kemper, & Borg, 2013). In general, higher measurement bias also decreases measurement precision, which is equal to weaker indicator-construct relationships (i.e., factor loadings or slope

parameters). As a consequence, differences in ARS would entail non-invariant factor loading patterns for the content factors (e.g. Welkenhuysen-Gybels, Billiet, & Cambré, 2003). This could cause a lack of *metric* MI (i.e., lack of invariance of factor loadings or slope parameters).

Second, by definition ARS leads to inflated mean scores on items (or item intercepts) regardless of semantic direction of the item (pro-trait/regular or con-trait/negative) (Cheung & Rensvold, 2000; Kankaraš, Vermunt, & Moors, 2011). As a consequence, differences in ARS would entail non-invariant item intercepts (e.g. Cheung & Rensvold, 2000; though see Little, 2000). This could cause a lack of *scalar* (intercept) MI.

However, it has been shown that measurements can appear fully invariant in MI tests, though ARS leads to erratic construct level differences between respondent groups (see Little, 2000; Thomas, Abts, & Vander Weyden, 2014; Weijters, Schillewaert, & Geuens, 2008). Weijters et al., for instance, found idiosyncratic mean differences in an unbalanced attitude scale across survey modes that disappeared after controlling for different response style behavior. The reason for this is that if response styles affect all items to a similar degree, a test of MI in intercepts might not detect such a uniform bias (Little, 2000; Steinmetz, 2013), rather the latent means and variances of the constructs could be affected (see Little, 2000, p.215).

Given these concerns, two arguments stand out for further investigating MI and the issue of response bias: (I) response style behavior should be controlled in order to accurately conduct MI tests, because (II) controlling the response style should generally make construct measurements better comparable across respondents (see Little, 2000; Morren, Gelissen, & Vermunt, 2012; Thomas et al., 2014; Weijters et al., 2008; Welkenhuysen-Gybels et al., 2003).

The present study

This study investigates if and how the presence of ARS impacts the results of MI tests and, accordingly, results on instrument comparability. For this purpose I will address the following central research question: does variation in ARS affect conclusions that one draws from MI tests, including the level of MI achieved? In other words, if we neglect ARS bias, will tests about MI come to the same conclusion? For the empirical analyses I apply multiple-group factor analysis (MG-FA) that can accommodate a powerful method for controlling ARS as a latent factor or as *random intercept*, i.e., a response factor varying over individuals (Aichholzer, 2014; Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006).

I continue by describing the methods used to assess the impact of ARS in MI tests and the substantial conclusions made by these tests. In order to study the impact of ARS, this study investigates MI among different educational groups, the reason being that research has generally found higher levels of ARS and/or higher variance of ARS due to lower formal education or lower cognitive abilities (e.g. Rammstedt, Goldberg, & Borg, 2010; Rammstedt & Kemper,

2011; Rammstedt et al., 2013; though see Waiyavutti, Johnson, & Deary, 2012). The article concludes with a discussion of implications of the findings, further applications, as well as potential future research.

Assessing measurement invariance with multiple-group factor analysis

As already mentioned, testing MI of instruments means investigating whether observed scores (using indicators) equally relate to latent constructs (or factors) in different contexts, which can generally be tested with multiple-group factor analysis (MG-FA) (Jöreskog, 1971). Factor analysis conceives j continuous latent factors or constructs as the common cause of k (continuous) observed measures or items using a linear model. Using matrix notation to denote the model gives the $k \times 1$ vector of responses to all observed measures (items) $\mathbf{y}$, the $k \times 1$ vector of item intercepts $\boldsymbol{\tau}$, the $k \times j$ matrix $\boldsymbol{\Lambda}$ of factor loadings that relate measures to the $j \times 1$ vector of factor scores $\boldsymbol{\eta}$, and the $k \times 1$ vector of uniquenesses (or residuals) $\boldsymbol{\varepsilon}$. This gives

$$\mathbf{y}=\boldsymbol{\tau}+\boldsymbol{\Lambda}\boldsymbol{\eta}+\boldsymbol{\varepsilon} \quad \text{(Eq. 1)}$$

It is usually assumed that residual variables ε_k are mutually uncorrelated and factors η_j are uncorrelated with residuals, i.e., $\mathrm{Cov}(\varepsilon_k, \varepsilon_l) = \mathrm{Cov}(\eta_j, \varepsilon_k) = 0$ for $\varepsilon_k \neq \varepsilon_l$. The implied (expected) variance-covariance matrix $\boldsymbol{\Sigma}_y$ of the observed variables y_k is then given by $\boldsymbol{\Lambda}$ times the factor variance-covariance matrix $\boldsymbol{\Psi}$ and the transpose $\boldsymbol{\Lambda}^{\mathrm{T}}$ plus the matrix of unique (residual) variances in $\boldsymbol{\Theta}$. This gives

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\mathrm{T}}+\boldsymbol{\Theta} \quad \text{(Eq. 2)}$$

The aim of multiple-group MI testing is to assess the equality of these measurement parameters, i.e., factor loadings in $\boldsymbol{\Lambda}$, item intercepts in $\boldsymbol{\tau}$, or the variance of item uniquenesses in $\boldsymbol{\Theta}$, etc. in a number of observed groups. In doing so, the parameters are equated in a sequential manner to assess whether consecutive levels of MI are achieved across groups (Vandenberg & Lance, 2000).[2]

While until recently this kind of multiple-group modeling was only available for the *restricted* factor analysis model or confirmatory factor analysis (i.e., MG-CFA) (for this notion see Seva & Ferrando, 2000), it can also be applied to the more general *unrestricted* or exploratory factor analysis model (i.e., MG-EFA) that places no restrictions on the item-factor loading structure, using the exploratory structural equation modeling (ESEM) framework (Asparouhov & Muthén, 2009).[3] Using unrestricted/EFA in the multiple-group case can be useful, because restricted/CFA models for measures of complex individual traits often fail to fit the data (e.g. Aichholzer, 2014; Asparouhov & Muthén, 2009; Marsh, Morin, Parker, & Kaur, 2014; Seva & Ferrando, 2000).

2 This is also called the *forward approach* (sequential constraints). Another approach would be the *backward approach* where constraints are sequentially released.

3 Note that recent extensions for assessing MI include the idea of *exact* vs. *approximate* (Bayesian) MI (B. O. Muthén & Asparouhov, 2012), whereas this paper is exclusively concerned with the traditional or exact MI approach.

Controlling acquiescence bias in measurement invariance tests

This section looks at how to control and mitigate measurement error associated with ARS when using MG-FA for assessing MI. One basic difference in the various approaches is whether ARS in the target scale items is controlled by using separate and dedicated marker variables (e.g. Watson, 1992; Weijters et al., 2008) or whether ARS is directly inferred from the items at hand (for examples see Savalei & Falk, 2014). While the former (*indirect*) method has already been used for inclusion in MI analyses (e.g. Weijters et al., 2008), it requires a large amount of additional items that have the mere purpose of measuring one's general response style behavior. The latter (*direct*) method, which will be applied here, represents a suitable solution for modeling ARS where items measuring the response style and the substantive constructs are identical. However, direct methods require the scale to be semantically balanced in order to be able to identify ARS, whereas the indirect method can also be applied to unbalanced target scales (Watson, 1992).

Three direct methods for controlling ARS have been used so far: (a.) ex-post standardization by subtracting the mean response across items (ipsatization) has been suggested for correcting raw scores (e.g. Fischer, 2004; Rammstedt & Farmer, 2013). Ipsatized data have frequently been analyzed with methods such as PCA, but rarely in multiple-group applications as they require further computational effort to be fitted within multiple-group MI tests (Cheung & Chan, 2002). (b.) EFA procedures with target rotation are also said to detect ARS (e.g. Lorenzo-Seva & Rodríguez-Fornells, 2006), though this method has not been applied in the MG-FA context or within ESEM. (c.) The restricted random intercept (RI) factor analysis approach has been used to reflect individual differences in a latent ARS factor (hereafter: RI/ARS factor method) (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006). In a simulation study this method was found to be relatively robust to violations of the assumption that ARS affects all items consistently or when using partially balanced scales (Savalei & Falk, 2014). The RI/ARS factor method has already been successfully applied in the multiple-group context with CFA (Welkenhuysen-Gybels et al., 2003), though the authors did not consider the impact of ARS on different aspects or levels of MI. In what follows, the RI/ARS factor method will be applied in the context of testing MI of instruments with MG-FA.

Testing measurement invariance with random intercept MG-FA

Random intercept factor analysis. The RI/ARS factor method for controlling ARS is convenient as it only requires adding one additional factor/variable (see Figure 1). Moreover, a RI/ARS factor can be added to restricted/CFA (i.e., RI-CFA) baseline models (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006) as well as to unrestricted/EFA (i.e., RI-EFA) baseline models (Aichholzer, 2014).[4]

The RI/ARS factor α_i varies over individuals (i.e., random factor) and has a loading vector set to 1 for all items (tau-equivalence, defined in vector **1**). It is

4 Ultimately, in the single-factor case RI-CFA and RI-EFA are statistically identical.

also restricted to be orthogonal to content factors and residuals, which is needed for model identification, i.e., $\mathrm{Cov}(\alpha,\eta_j)=\mathrm{Cov}(\alpha,\varepsilon_k)=0$. This restriction implies that the ARS level is independent from the respondents' content factor scores.[5] Thus, by adding the RI/ARS factor (here: its variance φ) degrees of freedom are reduced by 1, which will inherently increase model fit as long as φ is significantly different from zero (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006).

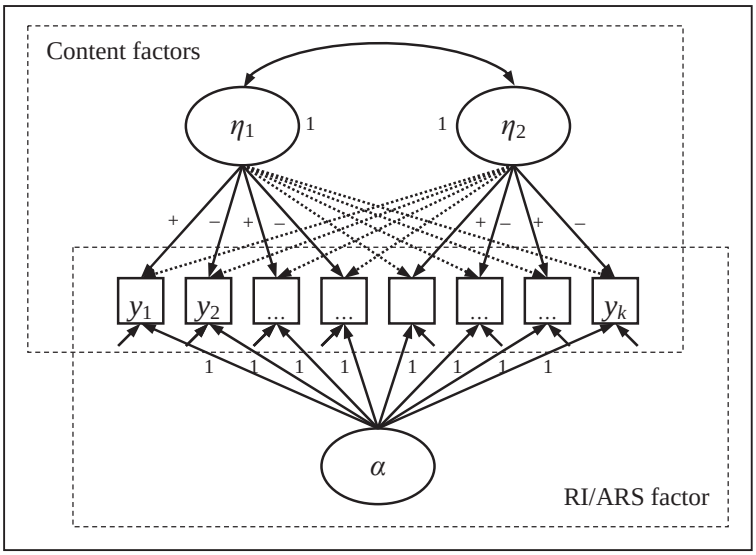

Figure 1. Graphical representation of random intercept factor analysis

(*Note.* Example for a two-factor model with semantically balanced scale in original coding. Residuals/uniquenesses ε_k are represented by arrows only)

The random intercept factor analysis (RI-FA) model in matrix form is

$$\mathbf{y}=\boldsymbol{\tau}+\boldsymbol{\Lambda}^*\boldsymbol{\eta}+\mathbf{1}\boldsymbol{\alpha}+\boldsymbol{\varepsilon} \qquad \text{(Eq. 3)}$$

whereas, in general, the implied $k \times k$ variance-covariance for RI-FA has the structure

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}^*\boldsymbol{\Psi}^*(\boldsymbol{\Lambda}^*)^{\mathrm{T}}+\mathbf{1}\varphi\mathbf{1}^{\mathrm{T}}+\boldsymbol{\Theta} \qquad \text{(Eq. 4)}$$

It shows that the indicators' (co)variance is now decomposed into common factor variance, systematic (co)variance due to ARS (φ), and unique (residual) factor variance.

When fitting a RI-CFA model with a restricted matrix $\boldsymbol{\Lambda}$, the usual identification rules in CFA apply (Jöreskog, 1969). When fitting RI-EFA, in order to be identified the condition must hold that the number of parameters to be estimated is equal or smaller than the number of empirical (co)variances

5 Note that uncorrelatedness of the ARS factor is, implicitly or explicitly, also the assumption of the other direct approaches (Savalei & Falk, 2014).

$$kj + \frac{[j(j+1)]}{2} + k - j^2 + 1 \leq \frac{[k(k+1)]}{2} \quad \text{(Eq. 5)}$$

where k is the number of indicators and j the number of factors. For instance, a 5-factor model ($j = 5$) with 10 items ($k = 10$) will be overidentified (d.f. = 4) and can thus be estimated with RI-EFA.

When fitting RI-EFA with an unrestricted matrix $\mathbf{\Lambda^*}$ for a hypothesized number of j content factors, the ESEM modeling framework (Asparouhov & Muthén, 2009) can be applied by using a rotation function $f(\mathbf{\Lambda})$, such as Varimax, Geomin, or Quartimin (see Sass & Schmitt, 2010), which gives $\mathbf{\Lambda^*}$ after rotation.[6] Note that the variance and latent mean of the RI/ARS factor are independent from the rotation function used (at equal number of content factors). In contrast, loadings in $\mathbf{\Lambda}$, the factor variance-covariance matrix $\mathbf{\Psi}$ as well as content factor means are contingent on the choice of the rotation function (denoted by an asterisk), while the intercept vector $\boldsymbol{\tau}$ and unique (residual) variances in $\mathbf{\Theta}$ are not (Asparouhov & Muthén, 2009, p.403).

Multiple-group random intercept factor analysis and testing MI. The general RI-FA model can readily be extended to a multiple-group model (MG-RI-FA). In the multiple-group case, all parameters are estimated separately for multiple groups ($g = 1, ..., G$) so that

$$\mathbf{y}_g = \boldsymbol{\tau}_g + \mathbf{\Lambda}_g^* \boldsymbol{\eta}_g + \mathbf{1}\boldsymbol{\alpha}_g + \boldsymbol{\varepsilon}_g \quad \text{(Eq. 6)}$$

Accordingly, the implied variance-covariance matrices are

$$\mathbf{\Sigma}_{yg} = \mathbf{\Lambda}_g^* \mathbf{\Psi}_g^* (\mathbf{\Lambda}_g^*)^{\mathrm{T}} + \mathbf{1}\boldsymbol{\varphi}_g \mathbf{1}^{\mathrm{T}} + \mathbf{\Theta}_g \quad \text{(Eq. 7)}$$

Again, the item-factor loading matrix $\mathbf{\Lambda}$ can be modelled to be restricted or unrestricted, using the rotated matrix $\mathbf{\Lambda^*}$ and multiple-group modeling capabilities in ESEM (Asparouhov & Muthén, 2009).

As already mentioned, testing MI means testing the equality of parameters in the factor analytic model across groups (Vandenberg & Lance, 2000). This can be done in a consecutive manner: first, *configural* MI or equality of the same baseline measurement model structure (i.e., j content factors underlying the indicators) is tested, which is generally a test of the similarity of the patterns of salient (target) loadings and non-salient loadings (secondary or cross-loadings) defining the constructs. Second, the unstandardized factor loading matrix (unrotated matrix in ESEM, see Asparouhov & Muthén, 2009, p.406) is constrained to equality, i.e., $\mathbf{\Lambda}_{g1} = \mathbf{\Lambda}_{g2} = \mathbf{\Lambda}$ to achieve *metric* MI. This is seen as a precondition for comparing construct correlations. Third, the indicator intercept parameters are constrained to equality, i.e., $\boldsymbol{\tau}_{g1} = \boldsymbol{\tau}_{g2} = \boldsymbol{\tau}$, to achieve *scalar* MI. This is seen as a precondition for comparing latent factor means, including the RI/ARS factor. Fourth, residual or uniqueness variances (denoted by the diagonal matrix $\mathbf{\Theta}$) are further equated, i.e., $\mathbf{\Theta}_{g1} = \mathbf{\Theta}_{g2} = \mathbf{\Theta}$, to achieve *uniqueness* MI, which means that constructs are measured identically. This allows comparison of

6 Variances of content factors are set to 1 for identification as in standard EFA.

explained variance for each indicator. Further, it has been suggested that some but not all parameters must be restricted in each step, i.e., allowing *partial* MI as the criterion when analyzing latent variables (Byrne, Shavelson, & Muthén, 1989).[7] Still, in order to adequately compare composite scores (summated scales) full scalar MI is required (Steinmetz, 2013).

If models are nested in such a stepwise manner, one can evaluate their equality by the chi-square difference test and/or changes in certain goodness-of-fit indices (ΔGOF). Since the χ^2-based MI test is known to be very sensitive to sample size and frequently results in rejection of MI, ΔGOF values are commonly used for judging levels of MI (Chen, 2007; Cheung & Rensvold, 2002).

Materials and methods

Samples

The present research is based on two samples.

The *ALLBUS* sample uses data from a large representative German population sample, the German General Social Survey (ALLBUS) 2008 (GESIS - Leibniz Institute for the Social Sciences, 2011) which, among others, administered the BFI-10 personality inventory (Rammstedt & John, 2007). The data are based on a random sample of the German adult population (n = 3469, age ≥ 18). Only participants who responded to all items of the BFI-10 and who provided educational information were included in the sample used for the analysis (n = 3118, age M = 50.3, SD = 17.6, 50.6% female). The BFI-10 was administered as part of the ISSP (International Social Survey Programme) module in a CASI (Computer Assisted Self-Interviewing) drop-off survey after a 45-min face-to-face interview.

The *ANES* sample uses data from a large U.S. representative population sample, the American National Election Study (ANES) Time Series Study 2012 (for details see ANES, 2014) which, among others, administered the 10-item TIPI personality inventory (Gosling, Rentfrow, & Swann, 2003). The data are based on a random sample of U.S. citizens (age ≥ 18 on election day). 93.2% or n = 5510 were re-interviewed in the post-election wave containing the TIPI. Only participants who responded to all items of the TIPI and who provided educational information were included in the sample used for the analysis (n = 5427, age M = 49.5, SD = 16.7, 51.3% female). The survey was administered in part by face-to-face interviews (35%) as well as via Web interviews (65%).

Measures

The BFI-10 (Rammstedt & John, 2007) and the TIPI (Gosling et al., 2003) are short and completely balanced 10-item scales for the Big Five personality traits: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. Each trait dimension is assessed by two semantically opposite measures (see Table 1 and Table 2 for the exact question wording). This semantic balance is important as it allows control and separation of ARS bias in measurement models.

Response categories for the BFI-10 are on a fully labeled Likert scale ranging from 1 (*applies completely*) to 5 (*does not apply at all*). Coefficient Alpha reliability estimates for the two items of each hypothesized dimension were .60 (E), .11 (A), .43 (C), .50 (S), .41 (O) in the *ALLBUS* sample.

---

7 Note however that ESEM does not allow for partial factor loading (metric) MI (see Marsh et al., 2014).

Table 1
Theoretical dimensions and items in the BFI-10

Domain	**Wording** I see myself as someone who…	**Direction**
Extraversion (E)	... is outgoing, sociable	pro-trait
	... is reserved	con-trait
Agreeableness (A)	... is generally trusting	pro-trait
	... tends to find fault with others	con-trait
Conscientiousness (C)	... does a thorough job	pro-trait
	... tends to be lazy	con-trait
Emotional Stability (S)	... is relaxed, handles stress well	pro-trait
	... gets nervous easily	con-trait
Openness to Experience (O)	... has an active imagination	pro-trait
	... has few artistic interests	con-trait

(Source: Rammstedt & John, 2007)

Response categories for the TIPI are on a fully labeled Likert scale ranging from 1 (*extremely poorly*) to 7 (*extremely well*). Coefficient Alpha reliability estimates for the two items were .45 (E), .28 (A), .52 (C), .52 (S), .38 (O) in the *ANES* sample.

Table 2
Theoretical dimensions and items in the TIPI

Domain	**Wording** Please mark how well the following pair of words describes you, even if one word describes you better than the other…	**Direction**
Extraversion (E)	... Extraverted, enthusiastic	pro-trait
	... Reserved, quiet	con-trait
Agreeableness (A)	... Sympathetic, warm	pro-trait
	... Critical, quarrelsome	con-trait
Conscientiousness (C)	... Dependable, self-disciplined.	pro-trait
	... Disorganized, careless	con-trait
Emotional Stability (S)	... Calm, emotionally stable	pro-trait
	... Anxious, easily upset	con-trait
Openness to Experience (O)	... Open to new experiences, complex	pro-trait
	... Conventional, uncreative	con-trait

(Source: ANES, 2014; Gosling et al., 2003)

Given the nature of measures (i.e., personality inventories) and response scales used (i.e., adjectives apply/do not apply or describe person well/poorly), the attribute associated with consistently endorsing the items resembles what Bentler et al. (1971) have called *acceptance acquiescence* or accepting characteristics as self-descriptive, rather than *agreement acquiescence* to general aphorisms.

Further note that both instruments have been tested with regard to validity and reliability in previous research (see Credé, Harms, Niehorster, & Gaye-Valentine, 2012; Gosling et al., 2003; Rammstedt & John, 2007). Nevertheless, there is still discussion revolving around the factorial structure and potential MI of these instruments. Several studies suggest that ARS should be adjusted in order to recover the theoretical five-factor structure of Big Five measures, while in heterogeneous samples this would make measurements more comparable (invariant) (e.g. Aichholzer, 2014; Rammstedt & Farmer, 2013; Rammstedt et al., 2010; Rammstedt et al., 2013). The present study therefore represents a replication and extension of previous work.

Method of analysis

In what follows, I first investigate initial model fit for the total (pooled) sample, using the least restrictive model, RI-EFA, as a starting point. Throughout the analyses the five-factor model of personality (Big Five) was hypothesized as measurement model (i.e., 5 latent factors), unless the model is extended by the RI/ARS factor (i.e., 5+1 latent factors). Each model allowed correlations between the latent personality variables. All analyses are carried out using the linear MLR estimator (Maximum Likelihood, robust standard errors for non-normality) in M*plus* Version 7 (Muthén & Muthén, 1998-2010). The oblique Geomin rotation criterion was selected for the EFA/ESEM analyses (see Asparouhov & Muthén, 2009; Sass & Schmitt, 2010).

Global model fit of each model is evaluated by the following goodness-of-fit (GOF) indices: Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Squared Residual (SRMR). The joint criteria of CFI > .90, RMSEA < .08, and SRMR < .08 are commonly regarded as good approximate fit and CFI > .95, RMSEA < .05, and SRMR < .05 as excellent approximate fit (on this issue see Marsh, Hau, & Wen, 2004). Further, using the same set of observed measures lower Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) values can be used as criteria for model selection.

For judging the level of MI, changes in goodness-of-fit indices (ΔGOF) are considered here (Chen, 2007; Cheung & Rensvold, 2002). Cheung and Rensvold (2002) suggested that a change of $\geq -.010$ in CFI is indicative of noninvariance of the more restricted model. Chen (2007, p.501) further specified that for "testing loading invariance, a change of $\geq -.010$ in CFI, supplemented by a change of $\geq .015$ in RMSEA or a change of $\geq .030$ in SRMR would indicate noninvariance; for testing intercept or residual invariance, a change of $\geq -.010$ in CFI, supplemented by a change of $\geq .015$ in RMSEA or a change of $\geq .010$ in SRMR would indicate noninvariance." While ΔCFI is sometimes regarded as the main criterion (Chen, 2007; Cheung & Rensvold, 2002), all reference values will be considered here.

Results

Evaluation of the baseline measurement model

Establishing an overall well-fitting baseline model is important as this model provides the basis for the joint estimation across groups. Tables 3a and 4a below therefore show the χ^2-test and global fit indices for different modeling strategies: (1.) RI-EFA, (2.) standard EFA, (3.) RI-CFA, and (4.) standard CFA. In some instances variances of residuals or factors had to be restricted (bounded) to be ≥ 0 for convergence after a Heywood Case in the initial solution (see Muthén & Muthén, 1998-2010, p.102).[8]

For both instruments the RI/ARS models showed an excellent and better fit by all criteria. In other words, omitting the RI/ARS factor or assuming zero variance of ARS results in a worse fit to the data. The models are nested so that they can be compared to the least restrictive model, RI-EFA (1.). A value of $\Delta CFI \geq -.010$ is considered as indicative of noninvariance of these nested models (Cheung & Rensvold, 2002). Further, AIC and BIC values also increased considerably as models were consecutively restricted, indicating that RI-EFA should be preferred. Further, the theoretical Big Five structure was supported by the RI-EFA model with Geomin rotation. Comparing the Geomin-rotated five-

8 Note that in this case modification indices, which usually provide the basis for possible model modifications (restricting/freeing parameters), are not computed in M*plus*.

factor loading matrix with an idealized perfect simple structure matrix, yielded Tucker's congruence coefficients of c = .93 and c = .90 for the BFI-10 and the TIPI, respectively (for detailed results see Tables A1 and A2 in the Appendix).

Testing measurement invariance and potential ARS bias

The next section will assess the level of MI in different educational groups. Analyzing different educational groups rests on the hypothesis that these groups differ with regard to the items' measurement properties. Different levels or variance in ARS can entail a lack of MI in self-reports across groups and, more generally, these might exhibit differential validity in responses (Rammstedt et al., 2010; Rammstedt & Kemper, 2011; Rammstedt et al., 2013). For simplicity of illustration three equally large groups were created, using the respondent's highest level of education: (a.) *Low* education (*ALLBUS*: lower secondary education or less n = 1187; *ANES*: up to high school credential n = 1906), (b.) *Intermediate* education (*ALLBUS*: intermediate secondary education n = 996; *ANES*: some post-high-school, no bachelor's n = 1818), and (c.) *High* education (*ALLBUS*: admission to tertiary education or completed university degree n = 935; *ANES*: Bachelor's or graduate degree n = 1703).

In the multiple-group analysis two variants were used: models without controlling for ARS (MG-FA) or taking into account ARS (MG-RI-FA). For this purpose the best fitting models RI-EFA (i.e., including the RI/ARS factor) and simple unrestricted EFA as specified for the total sample will be compared, respectively.[9]

Overall, the results in Tables 3b and 4b corroborate that the configural MI model with RI-EFA is to be preferred over standard EFA, as indicated by excellent goodness-of-fit values. This is a basic indication that ARS constitutes an additional factor that should be taken into account in all three subgroups. Following the criteria outlined above (ΔGOF), we now look at consecutive steps in testing MI (i.e., configural, metric, scalar, and uniqueness MI). First, the results suggest that regardless of controlling for ARS or not, in both samples full metric MI is supported as indicated by the ΔGOF values. Second, the evidence in support of full scalar MI is somewhat mixed. For the *ALLBUS* sample most indices point towards supporting full scalar MI (though CFI decreases by more than –.010), however, regardless of controlling ARS or not. For the *ANES* sample all indices support scalar MI in the model controlling ARS (RI-EFA), whereas there is no clear evidence for full scalar MI for the simpler EFA model. In other words, using very strict criteria, one might reject full scalar MI in one case (ARS not controlled), but not in the other (ARS controlled). Finally, uniqueness MI is tested and clearly supported for the *ALLBUS* sample and in part for the *ANES* data, regardless of the measurement model. However, with RI-EFA the full uniqueness MI model would be accepted for the three groups in both samples, as it has excellent fit to the data (*ALLBUS* sample: CFI = .970, RMSEA = .029, SRMR = .031; *ANES* sample: CFI

9 Substantial results on the impact of ARS in MI testing are quite similar when using the more restrictive models, MG-RI-CFA and MG-CFA, though worse in overall fit (see Tables S1 and S2 in the Supplemental Materials). However, the ΔCFI is generally larger for the more restrictive models in this case.

= .964, RMSEA = .039, SRMR = .044).[10] Especially for the *ANES* sample, the full uniqueness MI with standard EFA has only an acceptable goodness-of-fit (*ANES* sample: CFI = .918, RMSEA = .058, SRMR = .080).

Summarizing, irrespective of whether ARS has been controlled or not by using different modeling strategies, one would come to the conclusion that scalar and uniqueness MI of the instruments investigated in the three educational groups is supported. Thus, latent mean comparisons of the content personality factors, as well as ARS would be viable and the explained variance of the items can be compared meaningfully.

Table 3a

Summary of goodness-of-fit indices for different measurement models (ALLBUS, pooled sample)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
RI-EFA [a] (1)	35.26	4	<.01	.988	.050	.006	86783	87152					
EFA [b]	108.54	5	<.01	.961	.081	.017	86857	87220					
(2 vs. 1)									–.027	.031	.011	74	68
RI-CFA [c]	217.10	25	<.01	.929	.049	.034	88155	88397					
(3 vs. 1)									–.059	–.001	.028	1372	1245
CFA [d]	476.00	26	<.01	.833	.074	.045	88432	88668					
(4 vs. 1)									–.155	.024	.039	1649	1516

Note. 5(+1) factor solution for the BFI-10. Entries with grey shading indicate better fit values. *ALLBUS* 2008 data, $n = 3118$. [a] No further constraints. [b] Residual variances bounded to be > 0. [c] Residual variance of item A_{con} set to 0. [d] Residual variance of item A_{con} set to 0.

Table 3b

Summary of goodness-of-fit indices for testing measurement invariance (ALLBUS, educational groups)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
MG-RI-EFA[a]													
Configural (1)	18.53	12	.10	.998	.023	.006	86479	87585					
Metric	92.27	62	<.01	.989	.022	.020	86456	87260					
(2 vs. 1)									–.009	–.001	.014	–23	–325
Scalar	142.91	70	<.01	.972	.032	.024	86485	87240					
(3 vs. 2)									–.017	**.010**	**.004**	29	**–20**
Uniqueness	168.18	90	<.01	.970	.029	.031	86493	87128					
(4 vs. 3)									**–.002**	**–.003**	**.007**	8	**–112**
MG-EFA [b]													
Configural (1)	112.85	25	<.01	.967	.058	.024	86572	87599					
Metric	148.85	65	<.01	.968	.035	.025	86523	87309					
(2 vs. 1)									**.001**	**–.023**	**.001**	**–49**	**–290**
Scalar	189.63	75	<.01	.957	.038	.031	86541	87266					
(3 vs. 2)									–.011	**.003**	**.006**	18	**–43**
Uniqueness	227.56	96	<.01	.950	.036	.038	86545	87138					
(4 vs. 3)									**–.007**	**–.002**	**.007**	4	**–128**

Note. Entries with grey shading indicate better fit values. ΔGOF entries in boldface indicate support of more restrictive MI step according to criteria proposed by Chen (2007). *ALLBUS* 2008 data, $n_{Low} = 1906$/ $n_{Intermediate} = 1818$/ $n_{High} = 935$. [a] No further constraints. [b] Residual variances bounded to be > 0 in all models.

10 When using the backward approach (releasing constraints) this restrictive model would be accepted.

Table 4a
Summary of goodness-of-fit indices for different measurement models (ANES, pooled sample)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
RI-EFA [a)] (1)	33.22	6	<.01	.996	.029	.009	185391	185780					
EFA [b)]	245.56	8	<.01	.966	.074	.021	185635	186011					
(2 vs. 1)									–.030	.045	.012	244	231
RI-CFA [c)]	662.156	24	<.01	.909	.070	.048	186021	186292					
(3 vs. 1)									–.087	.041	.039	630	512
CFA [d)]	2718.49	28	<.01	.617	.133	.091	188493	188737					
(4 vs. 1)									–.379	.104	.082	3102	2957

Note. 5(+1) factor solution for the TIPI. Entries with grey shading indicate better fit values. *ANES* 2012 data, n = 5427. [a)] Residual variance of item O_{pro} and A_{pro} set to 0. [b)] Residual variance of item E_{pro}, S_{con} and C_{con} set to 0. [c)] No further constraints. [d)] Residual variance of item E_{pro}, A_{pro} and S_{pro} set to 0.

Table 4b
Summary of goodness-of-fit indices for testing measurement invariance (ANES, educational groups)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
MG-RI-EFA [a)]													
Configural (1)	39.23	18	<.01	.997	.026	.010	184422	185590					
Metric	160.21	68	<.01	.987	.027	.025	184461	185299					
(2 vs. 1)									**–.010**	**.001**	**.015**	39	**–291**
Scalar	207.77	76	<.01	.982	.031	.026	184496	185281					
(3 vs. 2)									**–.005**	**.004**	**.001**	35	**–18**
Uniqueness	348.81	92	<.01	.964	.039	.044	184625	185304					
(4 vs. 3)									–.018	**.008**	.018	129	23
MG-EFA [b)]													
Configural (1)	314.22	24	<.01	.960	.082	.024	184696	185825					
Metric	433.33	74	<.01	.950	.052	.035	184769	185568					
(2 vs. 1)									**–.010**	**–.030**	**.011**	73	**–257**
Scalar	538.64	84	<.01	.937	.055	.040	184861	185594					
(3 vs. 2)									–.013	**.003**	**.005**	92	26
Uniqueness	684.58	98	<.01	.918	.058	.080	185024	185665					
(4 vs. 3)									–.019	**.003**	.040	163	71

Note. Entries with grey shading indicate better fit values. ΔGOF entries in boldface indicate support of more restrictive MI step according to criteria proposed by Chen (2007). *ANES* 2012 data, n_{Low} = 1906/ $n_{Intermediate}$ = 1818/ n_{High} = 1703. [a)] Residual variance of item O_{pro} and A_{pro} set to 0 in all models. [b)] Residual variance of item E_{pro}, S_{con} and C_{con} set to 0 in all models.

The question arises, whether the subgroups investigated here actually differ in ARS and whether this has any impact on measures, since otherwise there would be no issues of bias in the MI tests. The scalar MI model is sufficient to compare the latent means of constructs, where the latent mean is fixed to 0 in one reference group (here: low education) (Vandenberg & Lance, 2000, p.57). We can further investigate information on the unstandardized variance of ARS and the average item variance explained by ARS from the uniqueness MI model.

Table 5 shows latent means and unstandardized variances of the RI/ARS factor for the *Low*, *Intermediate*, and *High* education group. Finally, the explained variance in items by ARS based on the uniqueness MI model was examined.

Summarizing, the variance of ARS, the level of ARS, and the explained variance by ARS are substantially lower among respondents with higher education. The hypothesis of equality in response style behavior (ARS) for different educational groups should thus be rejected. Larger impact and differences in ARS in the *ANES* sample might explain stronger divergence between the MI tests conducted with and without controlling ARS. In sum, differences in ARS between samples and groups might be due to the different survey modes used.

Table 5
Differences in acquiescence (RI/ARS factor) between educational groups

Data	Group (education)	Unstandardized latent mean	Unstandardized variance	Item variance explained
ALLBUS	Low	0 (fixed)	.046 (.006)	4.5%
(n = 3118)	Intermediate	–.077* (.017)	.035 (.006)	3.6%
	High	–.119* (.021)	.023 (.005)	2.5%
ANES	Low	0 (fixed)	.310 (.021)	14.2%
(n = 5427)	Intermediate	–.051* (.020)	.141 (.013)	7.2%
	High	–.073* (.020)	.072 (.009)	3.9%

Note. Estimates based on full uniqueness MI model. S.E. in parentheses, *mean difference significant at $p < .05$.

Discussion and conclusion

The overarching question of this study was if and how acquiescence response style (ARS) affects conclusions drawn from measurement invariance (MI) tests, including the level of MI achieved. The analyses were based on empirically testing MI of two short personality scales (BFI-10 and TIPI) in three different educational groups, using either standard multiple-group factor analysis or a model controlling for ARS by means of an additional random intercept (RI) factor. In fact, the analyses are identical to a test whether or not omitting the additional ARS factor leads to different conclusions with regard to MI of that scale.

Overall, the results of the present study suggest that the impact of different ARS on MI tests is negligible, even though it was shown that the groups significantly differed in terms of the level, variance, and thus the impact of ARS on measures (for similar results see Rammstedt et al., 2010). In other words, for the BFI-10 (*ALLBUS* data) one would conclude that the instrument is fully invariant either way and parameters, such as construct correlations with other variables and means can be meaningfully compared. So, is ARS for testing MI irrelevant?

The results are surprising in the light that previous literature suggests a biasing effect of ARS on several measurement parameters. Most research concludes that controlling ARS purifies item-factor structures so that they become more valid and comparable (i.e., they show metric or factor loading invariance), whereas "omitting a factor accounting for the acquiescent response bias leads to a biased assessment of the invariance of the loadings of the content factor across the groups under study" (Welkenhuysen-Gybels et al., 2003, p.720). There is, however, no indication whatsoever that assuming metric MI gives a

worse fitting model, regardless of whether ARS is controlled or not. Further, Cheung and Rensvold (2000) argued that lack of overall intercept invariance (lack of scalar MI) is indicative of different ARS across populations. Similarly, there is no clear indication of a lack of scalar MI, regardless of whether ARS is controlled or not. The results of this study rather corroborate the findings by Weijters et al. (2008) who find no indications of response style differences in the MI tests themselves, but eventually idiosyncratic construct level differences caused by the response style bias (also see Little, 2000).

Implications

The findings of this study suggest that, contrary to several observers, taking into account ARS or not does not strongly affect conclusions that one draws from MI tests. More precisely, in absolute terms, controlling for ARS improved the measurement models' fit, but controlling ARS has no clear impact for inferences on levels of MI achieved. In turn, this implies that the standard procedure of MI testing across groups, which is what Cheung and Rensvold (2000) suggested, does not seem to be an appropriate means for detecting ARS differences per se.

Latent variable approaches which can incorporate response style factors seem more promising in this regard, since "response styles must be measured independently of the other constructs under consideration, and the tests of the other constructs must be done while controlling and correcting for [response styles]" (Little, 2000, p.215). Along these lines, Thomas et al. (2014) also argue that "controlling RSs [response styles] while demonstrating MI should become a basic research requirement […] necessary in within-country and cross-cultural research". The point is that it remains a key issue in comparative (e.g. cross-cultural) research to disentangle both the equal/unequal *interpretation of question* content and equal/unequal response behavior (scale-usage), i.e., whether people differ in some general *response style*.

Limitations and future research

Some limitations should, nevertheless, be mentioned. A clear limitation is the use and selection of specific datasets, including specific samples, and the specific instruments, i.e., the number items and semantic balance of scales being analyzed. Furthermore, ARS was inferred from the items at hand, whereas other research has used separate marker items to control for ARS. Besides, the analyses were restricted to a specific estimation method and assumption, namely linear MLR estimation. Future studies might therefore use other data and/or simulations to investigate more general aspects with regard to the impact of response bias on MI tests to investigate various measurement conditions and various response styles. These aspects may include:

Respondent differences in response styles: An important factor that may impact measurement non-invariance of constructs could be the amount of variation in the mean level and/or variance of response styles across respondent groups. The larger these differences are, the larger the impact on the

measurements will be. It might be the case that ARS differences were simply too small to be detected in the different steps of testing MI. At the same time, this study did not consider other response styles such as extreme responding (ERS) or midpoint tendency (MRS) and their potential impact on measurement non-invariance (see Cheung & Rensvold, 2000; Morren et al., 2012; Thomas et al., 2014; Weijters et al., 2008).

Assessment of response styles: While this study only applied the *direct* method to detect and control ARS, future research might use *indirect* methods (marker variables) as well (see, for example, Watson, 1992; Weijters et al., 2008). These could be integrated in the general MG-FA framework and MI testing procedures as well. Another advantage of the indirect method can be resolving identification problems when style and content factors are to be correlated.

Items and constructs: The analyses were limited to a small set of items for measuring the substantive factors (2 × 5 items), though the scale was fully balanced. A larger number of items and/or sufficient heterogeneity of substantive constructs could be necessary in order to reliably identify the respondents' response style in the data. Future research could also look at situations where scales are not fully or not at all balanced, as previous research suggests that invariance in unbalanced scales shows up at the construct level, not at the item level (see Little, 2000; Thomas et al., 2014; Weijters et al., 2008). Again, the latter case would require additional marker items (Watson, 1992).

Estimation: Finally, research has argued that linear factor analysis for modeling ordered categorical Likert-type indicators could be problematic for detecting non-invariance in multiple-group MI tests (Kankaraš et al., 2011; Lubke & Muthén, 2004). Given these concerns, further applications could treat items as ordered categorical measurements. Further, besides factor analysis (or ESEM) other methods to investigate MI are available as well, such as item response theory (IRT) or latent class analysis (LCA) approaches (see Kankaraš & Moors, 2010; Kankaraš et al., 2011). At the same time, the present research links to other recent developments in MI analyses, such as using Bayesian specifications in MI analyses (Muthén & Asparouhov, 2012).

Given these limitations and suggestions, it would hence be of great interest to see future works studying the general conditions under which response styles can impact MI and under which conditions these could be detected by MI testing.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1-4.

ANES. (2014). *User's Guide and Codebook for the ANES 2012 Time Series Study*. The University of Michigan & Stanford University, Ann Arbor, MI & Palo Alto, CA.

Asparouhov, T., & Muthén, B. O. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397-438.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186-204.

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*(4), 542-562.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608-628.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*(2), 187-212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cheung, M. W.-L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*(1), 55-77.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874-888.

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality, 57*, 119-130.

Fischer, R. (2004). Standardization to Account for Cross-Cultural Response Bias: A Classification of Score Adjustment Procedures and Review of Research in JCCP. *Journal of Cross-Cultural Psychology, 35*(3), 263-282.

GESIS - Leibniz Institute for the Social Sciences. (2011). *ALLBUS/GGSS 2008 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2008) (ZA4600 Data file Version 2.0.0)*. GESIS Data Archive. Cologne.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. J. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology, 5*, 980.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Kankaraš, M., & Moors, G. (2010). Researching Measurement Equivalence in Cross-Cultural Studies. *Psihologija, 43*(2), 121-136.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. *Sociological Methods & Research, 40*(2), 279-310.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236.

Little, T. D. (2000). On the Comparability of Constructs in Cross-Cultural Research: A Critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31*(2), 213-219.

Lorenzo-Seva, U., & Rodríguez-Fornells, A. (2006). Acquiescent responding in balanced multidimensional scales and exploratory factor analysis. *Psychometrika, 71*(4), 769-777.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in

overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology, 10*, 85-110.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344-362.

McCrae, R. R., Herbst, J. H., & Costa, P. T., Jr. (2001). Effects of acquiescence on personality factor structures. In R. Riemann, F. M. Spinath, & F. Ostendorf (Eds.), *Personality and temperament: Genetics, evolution, and structure* (pp. 217-231). Berlin: Pabst.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E., & Meredith, W. (1992). Inferential Conditions in the Statistical Detection of Measurement Bias. *Applied Psychological Measurement, 16*(4), 389-402.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The Impact of Controlling for Extreme Responding on Measurement Equivalence in Cross-Cultural Research. *Methodology, 8*(4), 159-170.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313-335.

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wright (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control It. *Annual Review of Psychology, 63*, 539-569.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*(4), 1137-1145.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 44*(1), 53-61.

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*(1), 203-212.

Rammstedt, B., & Kemper, C. J. (2011). Measurement equivalence of the Big Five: Shedding further light on potential causes of the educational bias. *Journal of Research in Personality, 45*(1), 121-125.

Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five Personality Measurements for Acquiescence: An 18-Country Cross-Cultural Study. *European Journal of Personality, 27*(1), 71-81.

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 435*(1), 73-103.

Savalei, V., & Falk, C. F. (2014). Recovering Substantive Factor Loadings in the Presence of Acquiescence Bias: A Comparison of Three Approaches. *Multivariate Behavioral Research, 49*(5), 407-424.

Seva, U. L., & Ferrando, P. J. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica, 21*(3), 301-324.

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups: Is Partial Measurement Invariance Enough? *Methodology, 9*(1), 1-12.

Thomas, T. D., Abts, K., & Vander Weyden, P. (2014). Measurement Invariance, Response Styles, and Rural-Urban Measurement Comparability. *Journal of Cross-Cultural Psychology, 45*(7), 1011-1027.

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*(1), 4-70.

Waiyavutti, C., Johnson, W., & Deary, I. J. (2012). Do personality scale items function differently in people with high and low IQ? *Psychological Assessment, 24*(3), 545-555.

Watson, D. (1992). Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness. *Sociological Methods & Research, 21*(1), 52-88.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The Stability of Individual Response Styles. *Psychological Methods, 15*(1), 96-110.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409-422.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34*(6), 702-722.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of Extreme Response Style and Acquiescence Over 8 Years. *Assessment*. doi: 10.1177/1073191115583714

Appendix

Table A1
Fully standardized factor loadings for the BFI-10 using RI-EFA

Item	F1 (E)	F2 (A)	F3 (C)	F4 (S)	F5 (O)	RI/ARS
E_{con}	**–.72**	–.05	–.04	.02	.02	.16
A_{pro}	.19	**–.27**	.17	–.08	.15	.18
C_{pro}	–.02	.05	**–.52**	.02	.09	**.29**
S_{pro}	–.02	–.01	.03	**.80**	.00	.18
O_{pro}	.23	.05	–.04	.07	**.42**	**.20**
E_{pro}	**.65**	–.03	–.06	.04	.01	**.20**
A_{con}	.04	**.65**	.05	–.06	.03	.18
C_{con}	–.02	.08	**.69**	.05	.01	.18
S_{con}	–.16	.01	.27	**–.39**	–.01	.18
O_{con}	.03	.03	–.01	.03	**–.63**	.16

Note. Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (S), Openness to Experience (O). First and second largest loading per factor in boldface. *ALLBUS* 2008 data, full sample, $n = 3118$.

Table A2
Fully standardized factor loadings for the TIPI using RI-EFA

Item	F1 (E)	F2 (A)	F3 (C)	F4 (S)	F5 (O)	RI/ARS
E_{pro}	**.70**	.05	–.11	–.01	.01	.29
A_{con}	.15	**–.21**	.06	.36	.04	.27
C_{pro}	.04	–.01	**–.59**	–.04	.00	**.37**
S_{con}	–.10	.04	–.04	**.85**	.00	.26
O_{pro}	.00	–.01	.01	.00	**.95**	.32
E_{con}	**–.56**	.06	–.07	–.01	.01	.26
A_{pro}	.02	**.94**	.02	.00	.01	**.35**
C_{con}	.01	.02	**.77**	–.02	.01	.27
S_{pro}	–.06	.05	–.18	**–.48**	.05	.33
O_{con}	–.18	–.09	.22	.02	**–.22**	.28

Note. Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (S), Openness to Experience (O). First and second largest loading per factor in boldface. *ANES* 2012 data, full sample, $n = 5427$.

Supplemental Materials

Table S1

Summary of goodness-of-fit indices for testing measurement invariance (ALLBUS, educational groups)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
MG-RI-CFA[a)]													
Configural (1)	273.98	75	<.01	.925	.051	.038	86646	87371					
Metric	271.02	85	<.01	.930	.046	.039	86632	87297					
(2 vs. 1)									**.005**	**–.005**	**.001**	**–14**	**–74**
Scalar	401.89	93	<.01	.883	.057	.047	86761	87378					
(3 vs. 2)									–.047	**.011**	**.008**	129	81
Uniqueness	441.68	111	<.01	.875	.054	.050	86764	87272					
(4 vs. 3)									**–.008**	**–.003**	**.003**	3	**–106**
MG-CFA [b)]													
Configural (1)	542.68	78	<.01	.825	.076	.048	86922	87629					
Metric	523.83	88	<.01	.835	.069	.049	86910	87557					
(2 vs. 1)									**.010**	**–.007**	**.001**	**–12**	**–72**
Scalar	694.38	98	<.01	.775	.077	.056	87073	87659					
(3 vs. 2)									–.060	**.008**	**.007**	163	102
Uniqueness	734.27	116	<.01	.767	.072	.061	87081	87559					
(4 vs. 3)									**–.008**	**–.005**	**.005**	8	**–100**

Note. Entries with grey shading indicate better fit values. ΔGOF entries in boldface indicate support of more restrictive MI step according to criteria proposed by Chen (2007). *ALLBUS* 2008 data, n_{Low} = 1906/ $n_{Intermediate}$ = 1818/ n_{High} = 935. [a), b)] Residual variance of item A_{con} set to 0 in all models.

Table S2

Summary of goodness-of-fit indices for testing measurement invariance (ANES, educational groups)

Models (comparison)	MLR χ^2	d.f.	*p*	CFI	RMSEA	SRMR	AIC	BIC	ΔCFI	ΔRMSEA	ΔSRMR	ΔAIC	ΔBIC
MG-RI-CFA[a)]													
Configural (1)	788.24	73	<.01	.901	.074	.050	185072	185877					
Metric	758.78	83	<.01	.906	.067	.052	185092	185831					
(2 vs. 1)									**.005**	**–.007**	**.002**	20	**–46**
Scalar	842.44	91	<.01	.896	.068	.052	185160	185847					
(3 vs. 2)									**–.010**	**.001**	**.000**	68	16
Uniqueness	981.40	110	<.01	.879	.066	.064	185311	185872					
(4 vs. 3)									–.017	**–.002**	.012	151	25
MG-CFA [b)]													
Configural (1)	2650.58	84	<.01	.643	.130	.092	187379	188111					
Metric	2923.36	94	<.01	.607	.129	.102	187639	188306					
(2 vs. 1)									–.036	**–.001**	**.010**	260	195
Scalar	3170.03	104	<.01	.574	.128	.108	187856	188457					
(3 vs. 2)									–.033	**–.001**	**.006**	217	151
Uniqueness	3460.00	118	<.01	.535	.125	.106	188085	188593					
(4 vs. 3)									–.039	**–.003**	**–.002**	229	136

Note. Entries with grey shading indicate better fit values. ΔGOF entries in boldface indicate support of more restrictive MI step according to criteria proposed by Chen (2007). *ANES* 2012 data, n_{Low} = 1906/ $n_{Intermediate}$ = 1818/ n_{High} = 1703. [a)] Residual variance of item E_{pro} set to 0 in group *High*, set free again in Uniqueness model. [b)] Residual variance of item E_{pro}, A_{pro} and S_{pro} set to 0 in all groups and models.