Scale Evaluation with Exploratory Structural Equation Modeling and Simultaneous Adjustment for Acquiescence Bias

Julian Aichholzer[1]
Department of Methods in the Social Sciences, University of Vienna

Correspondence:

Julian Aichholzer
Department of Methods in the Social Sciences, University of Vienna
Rathausstraße 19/I, 1010 Vienna, Austria
e-mail address: julian.aichholzer@univie.ac.at
Phone: +43-1-4277-49909
Fax: +43-1-4277-9499

Draft prepared for
5th Conference of the European Survey Research Association (ESRA)
Ljubljana, July 2013

--- Please, do not cite or quote without authors' permission ---

Abstract: There is continued debate about the appropriateness of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) for scale evaluation and latent variable modeling more generally. This study seeks to combine the strengths of EFA, namely flexible evaluation of item-factor structures, while using one of the unique strengths of CFA, the simultaneous adjustment for common method bias in self-report survey data. Exploratory Structural Equation Modeling (ESEM) (Asparouhov & Muthén, 2009) is used to combine the EFA logic with CFA adjustment for individual acquiescence bias in semantically balanced scales. Two applications corroborate the usefulness of this model, using the short personality scale of Rammstedt and John (2007) as an example. The first application shows how the model can be used for general scale evaluation. The second application incorporates the model into testing measurement invariance of instruments while controlling for acquiescence across respondents. Finally, further applications of this model will be discussed.

Keywords: Exploratory Structural Equation Modeling; factor analysis; scale evaluation; acquiescence; measurement invariance

---

[1] Funding: This research is conducted under the auspices of the Austrian National Election Study (AUTNES), a National Research Network (NFN) sponsored by the Austrian Science Fund (FWF) (S10903-G11).

1. Introduction

The psychometric evaluation of self-report scales, that is, a collection of survey questions on the respondent's personality or attitudes, as well as their practical use is a vital task in psychology and the social sciences in general. In this realm the factor analytic technique has eventually become the most important statistical technique for scale evaluation and latent variable modeling. However, a large body of literature is concerned with the advantages or pitfalls of different approaches of the 'common factor model' (Thurstone, 1947), namely the differences between Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) (e.g. Asparouhov & Muthén, 2009; Hopwood & Donnellan, 2010; Marsh et al., 2010; Schmitt, 2011). A recent contribution has been set out to settle this cleavage: the Exploratory Structural Equation Modeling (ESEM) framework by Asparouhov and Muthén (2009), which now enables the researcher to combine features of EFA and CFA in a fruitful way.

The current study

This study argues that the newly developed ESEM approach allows for an important integration of statistical properties of EFA and CFA that has not been available, so far. In particular, the ESEM approach, on the one hand, makes it possible to use an important strength of EFA: the flexible investigation of complex indicator-factor structures, which often occur in the measurement of complex theoretical constructs, such as personality dimensions (e.g. Asparouhov & Muthén, 2009; Marsh et al., 2010). This property is one of the main reasons why EFA models usually perform better with regards to model fit to a hypothesized structure (Asparouhov & Muthén, 2009; Hopwood & Donnellan, 2010; Marsh et al., 2010; McCrae, Zonderman, Costa Jr., Bond, & Paunonen, 1996; Schmitt, 2011). On the other hand, ESEM allows using features of the CFA approach simultaneously. This is desirable since CFA has several properties and advantages over EFA that have, so far, been unique to the

former. Among other features, several CFA models have been developed to capture the impact of correlated measurement error or so called 'method effects' inherent in self-report instruments (see, for example, Podsakoff, MacKenzie, & Podsakoff, 2012). The rationale is that inclusion of method bias should result in higher validity of trait factors, on the one hand, and better model fit, on the other hand.

The idea to control for method bias, i.e. to separate 'trait' from 'method' variance in indicators, has already received large attention in early psychological research, most of which used CFA and the related multi-trait multi-method (MTMM) approach (Campbell & Fiske, 1959; see also Podsakoff et al., 2012; Pohl & Steyer, 2010). There is now wide evidence in the literature that so called method bias can arise from 'rater response styles, item characteristics, and aspects of the measurement context', which results systematic measurement bias and disturbs measurement validity (Podsakoff et al., 2012, p.542). Above all, literature on method bias in self-report surveys has tried to adjust for item wording effects associated with the respondent's acquiescence or agreeing response style to rating scales that use a Likert-type question format (e.g. Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Podsakoff et al., 2012; Rammstedt, Kemper, & Borg, 2013; Weijters & Baumgartner, 2012). In particular, this bias results in spurious correlations between questionnaire items (correlated errors). For this reason, personality or attitude scales often exhibit a blurred factorial structure, i.e. lack of convergent and discriminant validity, lack of evidence that a construct was unidimensional or occurrence of artificial factors (Podsakoff et al., 2012; Vautier & Pohl, 2009; Weijters & Baumgartner, 2012). Not being able to accurately control for the impact of acquiescence bias in the classical EFA model is thus seen as a vital drawback.

Contribution

This article presents an extended model of the ESEM approach by Asparouhov and Muthén (2009) that combines advantages of EFA, i.e. flexible evaluation of the relation

between indicators and latent factors, and CFA, i.e. statistical modeling of systematic acquiescence bias (or method bias) in semantically balanced scales. This model thus represents a useful tool in practical research on scale evaluation and latent variable modeling more generally.

The model presented here will be illustrated for two 'core tasks' in scale evaluation (see Brown, 2006) in the empirical section. First, it can be used for the purpose of 'classical' scale development and scale evaluation, i.e. assessment of factor loading structure and fit to a theoretical model (Application I). A hypothesized number of factors is tested with regards to indicator-factor structures (EFA part), while controlling for acquiescence bias in a confirmatory model (CFA part). Second, an example is provided that shows how the model presented here can readily be integrated into testing 'measurement invariance' of a scale (Meredith, 1993), i.e. whether the measurement model of constructs holds across respondent groups or measurement occasions (Application II). In doing so, one can include acquiescence response style as a separate factor while simultaneously testing the measurement invariance of an EFA factor structure (also see Asparouhov & Muthén, 2009; Marsh et al., 2010). Indeed, other scholars have stressed that control for acquiescence response style is crucial in the course of testing measurement invariance in comparative research as this response behavior distorts inferences on construct comparability or invariance (Cheung & Rensvold, 2000; Welkenhuysen-Gybels, Billiet, & Cambré, 2003).

The article, first, provides an introduction by contrasting EFA and CFA. After that, statistical remedies for the correction of method bias associated with acquiescence are presented. This discussion is used to synthesize the ESEM model presented in what follows. For all applications the M*plus* software (Muthén & Muthén, 1998-2012) is used, whereas the syntax is provided as online supplementary material for replication.[2] Both empirical

[2] See: To be added

applications examine the 10-item personality inventory (BFI-10) by Rammstedt and John (2007) that was administered in a large representative population sample. The article concludes with a discussion of implications and potential applications in further research.

2. Contrasting CFA and EFA

The common factor model

Most studies in psychology and other disciplines in the social sciences use EFA or CFA for scale evaluation and scale development, both of which are based on the so called 'common factor model' (Thurstone, 1947). Basically, the common factor model is a reflective measurement model, that is, it conceives latent traits (constructs or factors) as the common cause of indicators (measurements). It follows a linear model that can be written as

$$Y_{ik} = \tau_k + \Lambda \cdot \eta_{ij} + \varepsilon_{ik}$$

where Y_{ik} is a vector of k observed scores (measurements) of respondent i on the j^{th} trait factor η, Λ is the factor loading matrix with k rows (i.e. indicators) and j columns (i.e. factors), and ε_{ik} represents unique variance or random measurement error of the indicator. The item intercept τ_k represents the mean Y_k value where latent variables (factors) are zero. It is common in latent variable models to assume that residuals are uncorrelated with trait scores and all indicator residuals are uncorrelated, i.e. $Cov(\eta,\varepsilon)$=0 and $Cov(\varepsilon_k,\varepsilon_{k'})$=0. Thus, other sources of indicator covariation besides the common trait are neglected.

Basically, the equation depicted above can refer to a one-factor model or to several underlying factors. In the case of several factors, CFA follows a restricted (confirmatory) factor model that assumes perfect simple structure in Λ, where each indicator is assigned to only one particular factor and has zero cross-loadings to other factors. This is also called the 'independent cluster model' (ICM-CFA) that requires a clear measurement theory of each indicator. CFA allows to determine and to test factor loadings, between-trait (factor)

correlations given that the equation system is overidentified (sufficient degrees of freedom). For this reason CFA is still seen as the 'gold standard' in scale evaluation.

In contrast, within EFA the factor loadings in the matrix Λ allow a complex loading structure, i.e. loadings across multiple factors. Thus, one can investigate more complex patterns that otherwise (using CFA) may be overlooked. The peculiarity of EFA, however, is the indeterminate nature of the equation system (i.e. factor indeterminacy). Thus, factor loadings, between-trait correlations, and factor scores depend on a factor rotation criterion (e.g. Varimax, Quartimin, Geomin, etc.), i.e. a rotated matrix Λ^*, while each solution has the same fit to the data (for a discussion of rotation functions, see Asparouhov & Muthén, 2009; Sass & Schmitt, 2010). In classical EFA the item intercepts are not included, all measures are standardized, and only the rotated factor pattern matrix Λ^* is of interest. So, this equation reduces to

$$Y_{ik} = \Lambda^* \cdot \eta_{ij} + \varepsilon_{ik} .$$

Weighing advantages and disadvantages

Due to the common ground of CFA and EFA, a large body of the methodological literature has been devoted to the appropriateness of one method over the other (e.g. Schmitt, 2011), for instance, in studying the dimensionality of individual personality (Asparouhov & Muthén, 2009; Hopwood & Donnellan, 2010; Marsh et al., 2010; Schmitt, 2011). In what follows, important points are summarized in order to provide a basic understanding necessary to synthesize the model presented below.

As already mentioned, CFA is still seen as the gold standard in scale evaluation. It is based on clear measurement hypotheses (ICM-CFA) that form the basis for the measurement model, including its parameters to be tested. Testing a CFA model includes standard errors (significance) of factor loading parameters that are, however, not available after EFA in most software packages (Asparouhov & Muthén, 2009). Further, so far only the CFA approach included the possibility to test an instrument's 'measurement invariance' (Meredith, 1993).

This is usually done by consecutively restricting measurement parameters, i.e. factor structure, factor loadings, and item intercepts, to equality (e.g. Vandenberg & Lance, 2000). Testing measurement invariance was, so far, not possible for the EFA model (Asparouhov & Muthén, 2009). Another advantage of CFA over EFA is that factor scores are corrected for random measurement error or unreliability (see Marsh et al., 2010). Also, CFA allows for the inclusion of further 'artificial factors' that represent systematic measurement bias or method factors (for an overview, see Podsakoff et al., 2012; Pohl & Steyer, 2010), including acquiescence response style (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006). As a consequence, EFA factor scores would include both unreliability due to random measurement error and potential systematic bias, which is not desirable for using them in further analyses. Hence, the advantage of CFA is that factors are corrected for different sources of measurement error, which is useful in an extended structural model (SEM, structural equation model).

However, several issues call the appropriateness of ICM-CFA for scale evaluation into question. Several scholars have argued that CFA is actually not an appropriate means for the evaluation of complex individual traits, because of its strict rationale on measurement (Asparouhov & Muthén, 2009; Hopwood & Donnellan, 2010; Marsh et al., 2010; McCrae et al., 1996). More precisely, an indicator may seldom be a measure of one and only one complex trait. ICM-CFA would hence lead to biased (inflated) factor loadings and between-trait correlations (Asparouhov & Muthén, 2009). Critique against using ICM-CFA is also corroborated by several instances where a CFA model does not result in proper fit to hypothesized theoretical model. Rather EFA, using the same number of factors, performs better. This fact is, for instance, very well documented in the literature on personality scale assessment (see Hopwood & Donnellan, 2010; Marsh et al., 2010). Indeed, bad model fit in CFA led researchers to pursue more liberal cutoff criteria for 'acceptable' model fit (Hopwood & Donnellan, 2010). Equally, researchers sometimes undertake severe model

modifications, such as allowing idiosyncratic error correlations, that render the confirmatory approach meaningless (see, for this critique, Hopwood & Donnellan, 2010; Schmitt, 2011).

Summarizing, until now researchers had to weigh the suitability of using either EFA *or* CFA in the practical evaluation and application of attitude or personality scales. As has been argued, one of the main disadvantages of EFA is the inability to accurately adjust for measurement bias, while CFA is regarded as overly strict in its measurement assumptions. In what follows a synthesis is presented in order to overcome this cleavage.

3. Acquiescence bias and its adjustment

To guard against acquiescence bias in surveys, it has become established practice to use 'balanced scales', that is, scales that use 'pro-trait' (positively phrased) items and 'con-trait' (negatively phrased) items to measure a trait. The advantage of balanced scales is the possibility to identify the variance and impact of acquiescence response style on observed scores. Ex-post statistical adjustment is then used to partial out acquiescence variance, the purpose of which is the expected improvement regarding consistency between items with opposite wording, changes in item-factor relationships and factor correlations as well as higher validity in the relation between trait variables (factors) and other covariates.

Yet, there are several statistical remedies available in the literature that are said to detect and to correct for systematic acquiescence bias (see, for an overview, Podsakoff et al., 2012). However, approaches such as substracting the mean response by ipsatization (Rammstedt et al., 2013) or partialling the mean response from indicators (Ten Berge, 1999), 'correct' the observed scores in a previous step and then conduct the factor analysis. In this case acquiescence is not a testable part in the measurement model. Alternatively, EFA approaches (e.g. Lorenzo-Seva & Rodríguez-Fornells, 2006) bear on the disadvantage that the computed factor scores are not corrected for unreliability of indicators. In contrast, CFA approaches have tried to model acquiescence as an individual difference variable in a

confirmatory measurement model part (e.g. Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006), which allows to control for its impact on observed scores and test inter-individual differences like with any other latent factor. For this reason, this study will focus on the latter approach (Model: CFA+Acq).

By definition, different levels of acquiescence should result in different thresholds to use response categories that indicate agreement or affirmation (e.g. Cheung & Rensvold, 2000). Using CFA, the impact of an acquiescence style factor can be modeled as a linear equation, where acquiescence A_i represents an additive individual difference variable or 'random intercept' apart from the item intercept τ_k (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006). In contrast to the more generic term 'method bias', the response style model thus conceptualizes bias as 'stemming from well-defined behavioral tendencies' (Weijters, Schillewaert, & Geuens, 2008, p.410).

It is common to assume that acquiescence is a 'uniform response bias', i.e. equal impact on all items in a combined scale. Factor loadings on an 'artificial' acquiescence factor are therefore replaced by +1, which gives indicator scaling and leaves its variance to be freely estimated (Billiet & McClendon, 2000).[3] Factor loadings on substantial factors in Λ follow perfect simple structure (ICM-CFA):

$$Y_{ik} = \tau_k + \Lambda \cdot \eta_{ij} + 1 \cdot A_i + \varepsilon_{ik}.$$

The advantage of this model is that there is only one additional parameter to be estimated in the model (degrees of reduced by 1). The existence of acquiescence bias and improvement of the model is corroborated if standardized factor loadings on the style factor are significantly different form zero and the variance of the style factor is nonzero, but smaller than the variance of substantial traits (Billiet & McClendon, 2000). Further note that if

[3] Alternatively, one can use a different parameterization. If con-trait items have been reverse-coded to run in the same logical direction towards the trait (e.g. codes 1→5, 5→1), one must replace loadings on these items by –1 instead of +1 and, again, let the variance to be freely estimated. This is what will be done in the empirical application below.

acquiescence is neither correlated with trait factors nor the residual terms, i.e. $Cov(A,\varepsilon)=Cov(A,\eta)=0$ and maintaining $Cov(\eta,\varepsilon)=Cov(\varepsilon_k,\varepsilon_{k'})=0$, the model follows a simple additive decomposition of variance components in each indicator. Hence, it can be shown that acquiescence would remain a component of the indicator-specific residual component ε_{ik} and is erroneously considered random measurement error (Maydeu-Olivares & Coffman, 2006).

4. Developing a new model for scale evaluation

Using Exploratory Structural Equation Modeling (ESEM)

The sale evaluation model presented in the following builds on the Exploratory Structural Equation Modeling (ESEM) framework by Asparouhov and Muthén (2009). The point is that, like EFA, the basic ESEM model uses a factor pattern matrix that is freely estimated, i.e. assuming complex structure and using a rotation criterion. However, ESEM also comprises standard errors for the rotated factor loading solution, the possibility to test measurement invariance for an EFA model, the possibility to model residual correlations or inclusion of additional CFA model parts (Asparouhov & Muthén, 2009). Unlike classical EFA, item intercepts, factor means and factor slopes are also modelled for latent factors in ESEM. This allows testing these parameters, though using the logic of an EFA measurement model, i.e. complex structure. It is important to note, however, that factor means are affected by the rotation function used, while intercepts and unique item (residual) variances of indicators are not (Asparouhov & Muthén, 2009, p.403).

Using an a priori definition of the number of factors and having sufficient degrees of freedom, the ESEM model can be estimated and tested with regards to model fit.[4] Moreover, ESEM has been implemented in the widely used latent variable modeling software M*plus* (Muthén & Muthén, 1998-2012).

---

[4] For a detailed discussion on identification and estimation see the paper of Asparouhov and Muthén (2009).

While there are already some applications of ESEM in the are area of scale evaluation and measurement invariance testing (see, for example, Ellison & Levy, 2012; Lang, John, Lüdtke, Schupp, & Wagner, 2011; Marsh et al., 2010; Marsh, Nagengast, & Morin, in press; Pettersson, Turkheimer, Horn, & Menatti, 2012; Rosellini & Brown, 2011), previous studies have neglected the potential impact of systematic measurement bias in their data. This is a main drawback, since the overwhelming majority of instruments uses classical rating scales that are prone to the respondent's tendency to acquiescent responding (see, for example, Krosnick & Presser, 2010; Podsakoff et al., 2012; Weijters & Baumgartner, 2012). Because ESEM allows for an integration of EFA and CFA measurement model parts, it is shown how the CFA adjustment method presented above can be integrated into ESEM.

The combined model: ESEM with adjustment for an acquiescence factor

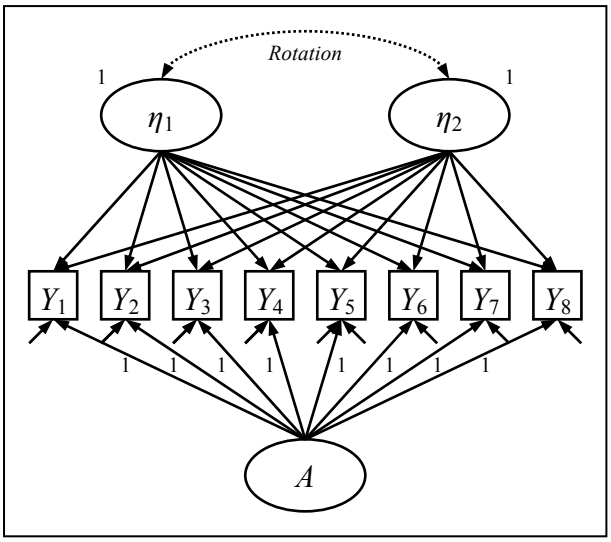
The model proposed here shows how ESEM can be used to assess item-factor loading structures in a flexible way, while making a simultaneous adjustment for systematic acquiescence bias with a restricted CFA model part (Model: ESEM+Acq). The general structure uses the same notation as the acquiescence factor model by Billiet and McClendon (2000) or Maydeu-Olivares and Coffman (2006)

$$Y_{ik} = \tau_k + \Lambda^* \cdot \eta_{ij} + 1 \cdot A_i + \varepsilon_{ik}.$$

However, while the acquiescence factor is based on a restricted CFA model, a complex factor pattern loading matrix Λ* can be examined for a hypothesized number of (substantial) factors, using a rotation criterion (see the Appendix for estimated variance and covariance in the model). That is, an EFA structure is examined, while controlling for acquiescence response behavior in a confirmatory fashion. Variances of the exploratory factors are fixed to 1 for identification purposes. All loadings on the acquiescence factor are restricted to +1 (or –1 otherwise, provided that an indicator has been reverse-coded first) (see Figure 1). Estimation of the acquiescence factor, i.e. factor loadings, its variance and latent mean, is independent from the EFA model part as it is a restricted confirmatory model. Note,

however, that factor means and factor slopes of exploratory (substantial) factors can be affected by rotation. Finally, all latent factors can then be used in further analyses when using ESEM.

Figure 1: Graphical representation of the ESEM+Acq measurement model



(Note: Residual variances ε_k are represented by arrows only)

5. Application

Sample and measures

In order to illustrate the ESEM+Acq model, recent data from the Austrian National Election Study (AUTNES 2013) pre-election survey are used. Participants come from a large representative sample of the Austrian eligible population aged 16 and above.[5] Households were chosen at random using an address-based stratified sample and then a random respondent was selected within each household. Interviews were conducted face-to-face (in-person) using computer assisted personal interviewing (CAPI). This survey provides a suitable application, since acquiescence is assumed to be more important in face-to-face than in self-administered interviews without an interviewer (Krosnick & Presser, 2010).

[5] Note: At the time of writing this paper, the survey field work was not completed yet. The full sample (approx. n=3,000) will be used in the final version of this paper.

The AUTNES survey has implemented the short 10-item Big Five personality Inventory (BFI-10) in German language (Rammstedt & John, 2007) that builds on the five-factor model of personality: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience (McCrae & Costa Jr., 2008). The scale has already been tested with regards to validity and reliability in previous research (Credé, Harms, Niehorster, & Gaye-Valentine, 2012; Rammstedt & John, 2007). However, as short scales go hand in hand with an inevitable reduction of construct validity there is still discussion revolving around the factorial structure of that instrument. Furthermore, several studies suggest the need to adjust for acquiescence in order to verify the five-factor personality structure (e.g. Rammstedt, Goldberg, & Borg, 2010; Rammstedt et al., 2013). Meanwhile, it has been argued that the factorial structure also varies in populations with different levels of acquiescence response style, such as different educational groups in a sample (Rammstedt et al., 2010). As such this study works both as a replication and extension of previous work.

The BFI-10 uses a completely balanced scale for the five personality dimensions (see Table 1). Response categories were on a 5-point Likert scale (1=’applies completely’, 5=’does not apply at all’) which permits estimation with linear structural equation models (see, for example, Maydeu-Olivares & Coffman, 2006). Coefficient alpha estimates for each hypothesized dimension were .64 (E), .17 (A), .47 (C), .44 (S), and .56 (O).

Table 1: Summary of theoretical dimensions and items in the BFI-10

Domain	Wording I see myself as someone who…	Direction	Label	Mean	S.D.
Extraversion (E)	... is outgoing, sociable	pro-trait	E_{pro}	2.30	.96
	... is reserved	con-trait	E_{con}	3.08	1.09
Agreeableness (A)	... is generally trusting	pro-trait	A_{pro}	2.64	1.06
	... tends to find fault with others	con-trait	A_{con}	3.20	1.02
Conscientiousness (C)	... does a thorough job	pro-trait	C_{pro}	1.72	.74
	... tends to be lazy	con-trait	C_{con}	3.74	1.14
Emotional Stability (S)	... is relaxed, handles stress well	pro-trait	S_{pro}	2.68	1.01
	... gets nervous easily	con-trait	S_{con}	3.56	1.02
Openness to Experience (O)	... has an active imagination	pro-trait	O_{pro}	2.33	.99
	... has few artistic interests	con-trait	O_{con}	3.21	1.27

Data: AUTNES 2013, n=776.

Analysis

In what follows, two applications are presented for the ESEM+Acq model: application I shows how the model can be used for general scale evaluation, while application II incorporates the model into testing measurement invariance in subpopulations. All analyses are carried out using linear ESEM with M*plus* Version 7 (Muthén & Muthén, 1998-2012), listwise deletion of missing values, and the MLR estimator (Maximum Likelihood, robust standard errors for nonnormality), which yields Satorra-Bentler chi-square values (Satorra & Bentler, 2001).

Throughout the analyses the basic five-factor model of personality is defined, unless the acquiescence factor is included so that the model is extended (i.e. 5+1 factors). Global model fit of each model is evaluated by significance of the χ^2-test, the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). In SEM models cutoff criteria CFI>.90, TLI>.90, RMSEA<.08 are commonly regarded as good fit and CFI>.95, TLI>.95, RMSEA<.05 as excellent fit (Marsh, Hau, & Wen, 2004). For nested models a lower Akaike Information Criterion (AIC) value indicates relative better fit. Comparisons of model fit with further restrictions on the model structure are conducted via the corrected chi-square difference test (Satorra & Bentler, 2001).

Application I: General scale evaluation

The following table (Table 2) shows global fit statistics for different modeling strategies: (I.) the CFA model, (II.) the EFA model, (III.) the CFA+Acq model, and (IV.) the ESEM+Acq model. However, except from the ESEM+Acq model all other models resulted in improper solutions, i.e. a so called 'Heywood Case' (H.C.) where estimation results yield standardized factor loadings greater than 1 or negative variance estimates, and models had to be further restricted to converge at all. This is an important indication of structural misspecifications for all other models. In comparison, the ESEM+Acq model for the BFI-10 had no estimation problems and showed an excellent fit to the data.[6]

Table 2: Summary of goodness-of-fit statistics for different measurement models

Model	MLR χ^2	d.f.	p	CFI	TLI	RMSEA	AIC
I. CFA[a]	84.27	25	<.01	.91	.83	.06	21724
II. EFA (or simple ESEM)[a]	10.61	5	.06	.99	.92	.04	21679
III. CFA+Acq[a]	72.71	24	<.01	.92	.86	.05	21713
IV. ESEM+Acq	3.02	4	.55	1.00	1.02	.00	21671

Note: [a]Heywood Case; results represent 5(+1) factor solution for the BFI-10; n=776.

In order to evaluate the factor loading structure (Table 3) all con-trait items have been reverse-coded first, so that loadings for pro-trait and con-trait items run in the same direction. Using this parameterization, factor loadings on the acquiescence factor follow a restricted +1/–1 pattern. For the EFA/ESEM model part the oblique Quartimin rotation solution was used, because it shows somewhat lower loadings on non-target factors and works well to approximate simple structure (Sass & Schmitt, 2010).

In general, the ESEM+Acq model supports the five-factor personality structure for the BFI-10. When looking at factor loadings in more detail, one can see that the S_{pro} item has several cross-loadings. Hence, CFA would have overlooked these minor cross-loadings. In

[6] Note: The findings on model fit and Heywood Cases can be replicated for recent data of the German Social Survey (ALLBUS 2008), which has been the basis for previous studies on the BFI-10 (e.g. Rammstedt, Goldberg, & Borg, 2010).

general, the findings also corroborate previous studies on the BFI-10 that find relatively weaker reliability for the Agreeableness indicators (Credé et al., 2012; Rammstedt & John, 2007) that, apparently, result from significant cross-loadings and lower common variance. The relatively stronger acquiescence factor loading on the C_{pro} item confirms that a good portion of agreement in this item is due to common bias. Finally, the results support the existence of a common systematic acquiescence bias to pro-trait and con-trait questions in the BFI-10. All factor loadings are significant and the factor variance is non-zero ($p<.01$).

Table 3: Summary of factor loadings from the ESEM+Acq measurement model

	E	A	C	S	O	Acq
E_{pro}	**.64***	.08	.05	-.02	-.02	.14*
E_{con}	**.76***	-.04	-.02	.05	.01	-.13*
A_{pro}	.15	**.43***	-.01	-.12	.09*	.13*
A_{con}	-.17	**.35**	.14*	.13	-.02	-.13*
C_{pro}	-.05	-.06	**.36***	.10	.11	.18*
C_{con}	.01	.01	**.95***	.00	-.01	-.12*
S_{pro}	-.07	.24	-.11*	**.45**	-.09*	.13*
S_{con}	.05	-.04	.04	**.77***	.03	-.13*
O_{pro}	.10	-.05	.06	.01	**.52***	.14*
O_{con}	-.03	.03	-.03	.01	**.79***	-.11*

Note: Quartimin rotation, factor loadings >|.30| are set in bold, *statistically significant at $p<.05$, n=776.

Application II: Measurement invariance

Next, an example illustrates the ESEM+Acq model for investigating measurement invariance. That is, the researcher tests whether observed scores in indicators measure the same latent constructs (factors) and equally relate to those constructs in different contexts (Meredith, 1993). Using the model presented here, one can include acquiescence response style as a separate factor while testing the measurement invariance of substantial constructs that are based on EFA factors, i.e. allowing cross-loadings. This is in line with other scholars who have stressed that control of differential response behavior is a crucial methodological prerequisite for making inferences on construct validity in comparative research (e.g. van de Vijver, 1998; Welkenhuysen-Gybels et al., 2003).

This study examines the hypothesis that self-report instruments sometimes do not permit measurement invariance in different educational groups (e.g. Rammstedt et al., 2010; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). Specifically, theories on survey responding suggest that people with lower education or cognitive capabilities are, in general, more prone to exhibit acquiescent responding (Krosnick & Presser, 2010). Different acquiescent behavior, in turn, results in blurred factorial structure and biased or unequal item intercepts between groups (Cheung & Rensvold, 2000).

The equality of measurement parameters, i.e. factor loadings and item intercepts, can be tested by means of multiple-group CFA (Vandenberg & Lance, 2000), while this procedure is now available for EFA/ESEM models too (Asparouhov & Muthén, 2009). Usually, the first step in testing measurement invariance is to establish 'configural invariance', that is, the same basic measurement model should hold (here: the 5+1 factor structure). This is seen as a core premise for the following steps. First, the measurement scale or the meaning of scale points is supposed to be identical across different respondents (='metric invariance'). For this purpose the factor loading matrix is constrained to equality, i.e. $\Lambda_{g1}=\Lambda_{g2}=\Lambda_G$ (for ESEM: Asparouhov & Muthén, 2009, p.406). Second, the scale origin of an indicator (the intercept) should be identical ('scalar invariance'). When testing scalar invariance, item intercepts are constrained to equality, i.e. $\tau_{g1}=\tau_{g2}=\tau_G$. The latter is seen as a precondition for comparing latent factor means afterwards. If models are nested in such a way, one can compare equality of model fit with the chi-square difference test (Satorra & Bentler, 2001).

For simplicity of illustration, we use two groups based on their highest level of education suitable for the Austrian context: (1.) individuals with a school-leaving certificate from upper secondary level and admission to higher (tertiary) education (n=263) or (2.) people with formal education beyond that level (n=511). Results in Table 4 support that the measurement instrument is scalar invariant in the two educational groups. Invariance of the more restricted models is supported as indicated by a non-significant scaled chi-square

difference test (Satorra & Bentler, 2001). Also, CFI does not decrease by more than .010 and RMSEA does not increase by more than .015 (Chen, 2007). Note that that these results do not depend on the rotation criterion used.

Table 4: Summary of goodness-of-fit statistics for testing measurement invariance

Model	MLR χ^2	d.f.	p	CFI	TLI	RMSEA	AIC	p($\Delta\chi^2$)
Configural invariance	4.34	8	.82	1.00	1.07	.00	21615	
Metric invariance	32.79	34	.53	1.00	1.01	.00	21591	.28
Scalar invariance	36.62	37	.49	1.00	1.00	.00	21533	.49

Note: All results represent solution for the ESEM+Acq model for different educational groups.

Next, we move on to a comparison of latent factor means. Table 5 shows latent mean differences for the substantial factors, i.e. personality dimensions, and the acquiescence style factor.[7] In order to identify the multi-group ESEM model, latent means were set to zero in one reference group (category: lower education). The results suggest that higher educated individuals report to be less conscientious but are considerably more open to experience. These educational differences in personality self-reports are largely in line with previous findings from a German population sample (Rammstedt, 2007). As expected, acquiescence response style is somewhat lower among higher educated individuals, though the difference is not statistically significant in this sample (p=0.29).

Table 5: Summary of latent mean differences between educational groups

Factor	Diff.	S.E.
Extraversion	-.01	.11
Agreeableness	.03	.09
Conscientiousness	-.36*	.11
Emotional Stability	.03	.09
Openness to Experience	.82*	.11
Acquiescence factor	-.04	.03

Note: Lower education is reference category (latent mean=0), *difference significant at p<.01, n=776.

[7] Note: This procedure equals regressing the latent factors on a manifest variable indicating the groups (see Lee, Little, & Preacher, 2011). Different factor rotation criteria (Geomin, CF-Varimax) had a negligible impact on latent mean results.

6. Discussion

Implications

The model presented in this study has important implications for research on scale evaluation, construct validation and latent variable modeling with self-report survey data more generally. This study sought to add a useful tool to the repertory of (1.) general scale evaluation, (2.) the examination of measurement invariance of an instrument, and (3.) further analyses with obtained factor scores from the measurement model. For this purpose flexible properties of EFA, detailed evaluation of the factor loading structure, have been combined with strengths of CFA, namely modeling systematic measurement error or acquiescence bias by means of the ESEM framework. Apparently, this approach is more flexible then the classical EFA *or* CFA approach. Still, as Marsh et al. (2010) have argued, ESEM might be a step prior to CFA. If a CFA model fits the data equally well, it should be preferred because of the possibility to conduct more rigorous tests on factor correlations, for instance.

Limitations

Some limitations should be mentioned, nevertheless. First, for illustrative purposes this study used a relatively short scale compared to other applications, especially in the psychological literature. Still, the model presented here can easily be extended. One could additionally integrate combinations of correlated uniquenesses or bi-factor models to capture sub-facets of larger construct (personality) dimensions (e.g. Marsh et al., 2010). Also, it is possible to extend the model to a situation where the indicators are not perfectly balanced. However, it is obvious that construct validity of the acquiescence factor increases with the number of items and the extent to which they are balanced. Second, the acquiescence style factor, as defined here, is specific to the BFI-10 scale and is not considered ‘pure acquiescence’. On the one hand, one should test whether acquiescence to questions in one scale generalizes to other scales in a questionnaire. On the other hand, one could add external indicators (marker variables) to further elaborate on its content validity (for that approach, see

Weijters et al., 2008; Williams, Hartman, & Cavazotte, 2010). Third, it has also been argued that linear modeling for categorical Likert-type indicators is not appropriate, especially for multi-group comparisons (Lubke & Muthén, 2004). Given these concerns, the ESEM model could be estimated under the assumption of categorical measures using a WLS estimator (Asparouhov & Muthén, 2009). Finally, it is still open to debate whether acquiescence bias, or more generally, method bias should be considered having equal impact on all measures and no correlation to the substantial traits (see Podsakoff et al., 2012). However, aside from theoretical issues, these questions essentially bear on the viability of parameter identification in a model.

Further research

In general, future research may elaborate on the usefulness of the ESEM+Acq model when testing measurement invariance. For instance, it may prove useful in cross-cultural research where one would like to test equivalence of the structure of theoretical constructs. Here it remains a key issue to disentangle both the interpretation of question content *and* different response behavior (scale-usage) that also affects factor analytic results (see, for this issue, van de Vijver, 1998; Welkenhuysen-Gybels et al., 2003). Equally, the model presented here can be used for multi-occasion modeling of panel data. Since the acquiescence factor is part of the measurement model, one can assess both its stability as well as its volatile impact on self-reports over time (for related issues, see Billiet & Davidov, 2008; Soto, John, Gosling, & Potter, 2008).

The main advantage of the ESEM approach eventually is that latent factor scores (means) and item intercepts are part of the model and can be used together with covariates (Asparouhov & Muthén, 2009; Muthén & Muthén, 1998-2012). Thus, the model can readily be integrated into the Multiple Indicators Multiple Causes (MIMIC) framework (Muthén, 1989). Using MIMIC, latent factors and a specific indicator are regressed on covariates (e.g. education). In doing so, one can test so called differential item functioning (DIF) or inequality

of item intercepts (Lee, Little, & Preacher, 2011) while controlling for the level of a target trait, cross-loadings to other factors, and acquiescence response behavior.

Finally, it should be mentioned that the rationale of the ESEM+Acq model can be apllied to all balanced scales in survey research, especially when measuring several theoretically distinct constructs. Hence, it would be of great interest to see a wider application in future work of different disciplines.

Appendix

The following restrictions are maintained for the ESEM+Acq model:

$$Cov(\eta, \varepsilon) = Cov(\eta, A) = Cov(\varepsilon, A) = Cov(\varepsilon_k, \varepsilon_{k'}) = 0\,.$$

So, the ESEM model with adjustment for acquiescence results in the following variance decomposition of indicators (see Brown, 2006):

$$Var(Y_k) = \lambda_{k1}^2 \cdot Var(\eta_1) + \lambda_{k2}^2 \cdot Var(\eta_2) + ... + \lambda_{kj}^2 \cdot Var(\eta_j) + \lambda_{kA}^2 \cdot Var(A) + Var(\varepsilon_k)$$

and using restrictions for identification

$$Var(Y_k) = \lambda_{k1}^2 + \lambda_{k2}^2 + ... + \lambda_{kj}^2 + Var(A) + Var(\varepsilon_k).$$

The covariance between two indicators can then be expressed as

$$Cov(Y_k Y_{k'}) = \lambda_{k1} \cdot Var(\eta_1) \cdot \lambda_{k'1} + \lambda_{k2} \cdot Var(\eta_2) \cdot \lambda_{k'2} + ... + \lambda_{kj} \cdot Var(\eta_j) \cdot \lambda_{k'j} + \lambda_{kA} \cdot Var(A) \cdot \lambda_{k'A}$$

and, using the restrictions from above

$$Cov(Y_k Y_{k'}) = \lambda_{k1} \cdot \lambda_{k'1} + \lambda_{k2} \cdot \lambda_{k'2} + ... + \lambda_{kj} \cdot \lambda_{k'j} + Var(A).$$

Thus, it can be shown that the model takes into account the information on all item cross-loadings as well as variance and covariance due to acquiescent responding.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling*, *16*, 397-438.

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*, 542-562.

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling*, *7*, 608-628.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling*, *14*, 464-504.

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*, 187-212.

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*, 874-888.

Ellison, W. D., & Levy, K. N. (2012). Factor Structure of the Primary Scales of the Inventory of Personality Organization in a Nonclinical Sample Using Exploratory Structural Equation Modeling. *Psychological Assessment*, *24*, 503-517.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332-346.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2 ed., pp. 263-313). Bingley Emerald.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: robust across survey methods except telephone interviewing. *Behavior Research Methods*, *43*, 548-567.

Lee, J., Little, T. D., & Preacher, K. J. (2011). Methodological issues in using structural equation models for testing differential item functioning. In E. Davidov, P. Schmidt & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 55-84). New York: Routledge.

Lorenzo-Seva, U., & Rodríguez-Fornells, A. (2006). Acquiescent Responding in Balanced Multidimensional Scales and Exploratory Factor Analysis. *Psychometrika*, *71*, 769-777.

Lubke, G. H., & Muthén, B. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. *Structural Equation Modeling*, *11*, 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*, 320-341.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U. (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. S. (in press). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects. *Developmental Psychology*, *Doi: 10.1037/a0026913*.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random Intercept Item Factor Analysis. *Psychological Methods*, *11*, 344-362.

McCrae, R. R., & Costa Jr., P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3 ed., pp. 159-181). New York: Guilford Press.

McCrae, R. R., Zonderman, A. B., Costa Jr., P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552-566.

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

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7 ed.). Los Angeles, CA: Muthén & Muthén.

Pettersson, E., Turkheimer, E., Horn, E. E., & Menatti, A. R. (2012). The General Factor of Personality and Evaluation. *European Journal of Personality*, *26*, 292-302.

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.

Pohl, S., & Steyer, R. (2010). Modeling Common Traits and Method Effects in Multitrait-Multimethod Analysis. *Multivariate Behavioral Research*, *45*, 45-72.

Rammstedt, B. (2007). The 10-Item Big Five Inventory: Norm Values and Investigation of Sociodemographic Effects Based on a German Population Representative Sample. *European Journal of Psychological Assessment*, *23*, 193-201.

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*, 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*, 203-212.

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*, 71-81.

Rosellini, A. J., & Brown, T. A. (2011). The NEO Five-Factor Inventory: Latent Structure and Relationships With Dimensions of Anxiety and Depressive Disorders in a Large Clinical Sample. *Assessment*, *18* 27-38.

Sass, D. A., & Schmitt, T. A. (2010). A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis. *Multivariate Behavioral Research*, *435*, 73-103.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514.

Schmitt, T. A. (2011). Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment*, *29*, 304-321.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The Developmental Psychometrics of Big Five Self-Reports: Acquiescence, Factor Structure, Coherence, and Differentiation From Ages 10 to 20. *Journal of Personality and Social Psychology*, *94*, 718-737.

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.

Ten Berge, J. M. F. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research*, *34*, 89-102.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: Univ. of Chicago Press.

van de Vijver, F. (1998). Towards a Theory of Bias and Equivalence. In J. A. Harkness (Ed.), *Cross-Cultural Survey Equivalence* (pp. 41-65). Mannheim: ZUMA.

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.

Vautier, S., & Pohl, S. (2009). Do Balanced Scales Assess Bipolar Constructs? The Case of the STAI Scales. *Psychological Assessment*, *21*, 187-193.

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*, 737-747.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 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* 702-722.

Williams, L. J., Hartman, N., & Cavazotte, F. (2010). Method Variance and Marker Variables: A Review and Comprehensive CFA Marker Technique. *Organizational Research Methods*, *13*, 477-514.