Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models

DL Oberski

Tilburg University, The Netherlands

ABSTRACT

Latent variable models can only be compared across groups when these groups exhibit measurement equivalence or “invariance”, since otherwise substantive differences may be confounded with measurement differences. This article suggests examining directly whether measurement differences present could confound substantive analyses, by examining the EPC-interest. The EPC-interest approximates the change in parameters of interest that can be expected when freeing cross-group invariance restrictions. Monte Carlo simulations suggest that the EPC-interest approximates these changes well. Three empirical applications show that the EPC-interest can help avoid two undesirable situations: first, it can prevent unnecessarily concluding that groups are incomparable, and second, it alerts the user when comparisons of interest may still be invalidated even when the invariance model appears to fit the data.

R code and data for the examples discussed in this article are provided in the electronic appendix (`http://hdl.handle.net/1902.1/21816`).

1. INTRODUCTION

Latent variable models are a common tool across the social sciences to model unobserved traits (Lord and Novick, 1968; Bollen, 1989; Bartholomew, Knott and Moustaki, 2011). In political science, for example, Jackman (2001) and Clinton, Jackman and Rivers (2004) applied ideal point models to roll calls in the U.S. Supreme Court, Senate, and House; Treier and Jackman (2008) and Armstrong (2011) discussed measurement models for the level of democracy based on the Polity and Freedom House indicators; and Davidov (2009) studied the measurement of national identity and constructive patriotism in the cross-national ISSP survey. The final goal of such analyses is often to compare estimated levels of the latent variable across groups; for instance, the level of democracy over time (Armstrong, 2011), the risk of civil war over levels of democracy (Treier and Jackman, 2008), or patriotism and nationalism over 34 countries (Davidov, 2009). Relationships between latent variables may also be compared – for example, the degree to which "human value priorities" predict attitudes against immigration across 19 European countries (Davidov et al., 2008).

Latent variables must be measured indirectly. Latent means and relationships may then only be compared across groups when group differences in these parameters of interest are unconfounded with group differences in measurement parameters (e.g. Steenkamp and Baumgartner, 1998; Oberski, 2012). One way to guarantee that differences of interest are not measurement artefacts is to establish that all (or some) measurement parameters are equal across groups. That is, to establish (partial) "measurement equivalence" or "invariance" (Meredith, 1993), often using structural equation models (SEM) for continuous variables. Chi-square difference testing (Steenkamp and Baumgartner, 1998; French and Finch, 2006), modification indices and expected parameter changes of the equality constraints (Byrne, Shavelson and Muthén, 1989;

	Misspecified invariance model fit	
	"Good" fit	"Bad" fit
Conclusions		
Unaffected by misspecification	(1) √	(2) Overparameterization or unnecessarily discarded item, group, or scale.
Affected by misspecification	(3) Non-invariance invalidates conclusions.	(4) √

Table 1: When the invariance model is misspecified, four situations may arise after fitting the invariance model.

Yoon and Millsap, 2007; Saris, Satorra and Van der Veld, 2009), as well as (change in) CFI, AGFI, RMSEA, and RMR (Hu and Bentler, 1998; Cheung and Rensvold, 2002; Chen, 2007) are then commonly used to assess invariance model fit (for reviews, see Millsap and Everson, 1993; Vandenberg and Lance, 2000; Schmitt and Kuljanin, 2008). A second way to rule out measurement artefacts as alternative explanations of substantive differences is the focus of this article: examining sensitivity, the likely impact of measurement differences on substantive comparisons of interest.

Focusing on equality of measurement parameters and not on whether measurement equality matters for conclusions of interest may lead to problematic situations when exact equality does not hold, as illustrated in Table 1. Columns in Table 1 represent whether invariance was rejected or not using any of the invariance evaluation procedures discussed, while rows represent the consequences for substantive comparisons of interest.

Situations (1) and (4) in Table 1 are unproblematic. In situation (1), the misspecifications present are inconsequential for the parameters of substantive interest, and the "good" model fit reflects this. Situation (1) is likely to occur when the misspecifications are small and the conclusions are not sensitive to the misspecifications in question. In situation (4), the misspecifications would have invalidated substantive comparisons if left undetected, but the model fit evaluation has correctly signaled their presence, which will tend to occur when misspecifications are large and the parameters of interest sensitive to them.

In contrast, situations (2) and (3) in Table 1 can seriously harm inference. While misspecifications may be present and even appear substantial, the parameters of interest are not necessarily sensitive to them (situation 2). An – arguably worse – situation occurs in situation (3). Here, fit indices signal a "close fit", but, due to large sensitivity, even "small" misspecifications substantially change parameters of interest. In this situation, even though an invariance model may appear to fit well, measurement artefacts still invalidate comparisons of substantive interest: precisely the problem that invariance testing was intended to prevent.

To prevent unnecessary information loss or a false sense of security when comparing groups using latent variables, invariance testing (columns of Table 1) should thus be supplemented with sensitivity analysis (rows of Table 1). This article suggests using the "EPC-interest" for that purpose: a measure of the expected change in the parameter of interest when freeing a particular equality restriction (Satorra, 1989; Bentler and Chou, 1992). Given parameters of interest to be compared across groups, the EPC-interest allows the researcher to establish more directly whether such a comparison is valid. Reanalyses of published examples show that both problematic situations (2) and (3) arise in practice, and how the EPC-interest can be used to prevent needless information loss and warn of otherwise unnoticed threats to cross-group comparisons.

The EPC-interest is different from the well-known "expected parameter change" (EPC) introduced to structural equation modeling by Saris, Satorra and Sörbom (1987). When considering what, hypothetically, can be expected to happen after freeing an equality constraint on a measurement parameter, the EPC estimates the expected change in that same measurement parameter, whereas the EPC-interest estimates the expected change in the parameter(s) of interest. It thus evaluates the sensitivity of the substantive model of interest to the invariance restrictions, and is similar in spirit to the approach for causal inference discussed by Imai and Yamamoto (2010, pp. 552-3). The EPC-interest approach also differs somewhat from the purely derivative-oriented approach to sensitivity analysis common in econometrics (Magnus and Vasnev, 2007, p. 168) and applied to SEM by Yuan, Marshall and Bentler (2003): both direction and magnitude of the misspecification are combined into the same measure here. A contingent hypothesis test of no change in the parameters of interest is, however, equivalent to classic econometric specification tests (Yuan, Marshall and Bentler, 2003; Hausman, 1978). Changes in parameters of interest for specific combinations of measurement and structural models were derived by Millsap (1997), Millsap and Kwok (2004), Millsap (2007), and Meuleman (2012). The EPC-interest can be seen as a general method of obtaining such results applicable to all structural equation models with equality-constrained measurement parameters.

The following section defines the EPC-interest for structural equation models with equality constraints. Subsequently, a simulation study evaluates the finite sample performance of the EPC-interest as an estimate of the shift in parameters of interest when freeing misspecified equality restrictions, as well as its robustness to misspecification in the alternative model. Sections 4, 5 and 6 demonstrate the use of the EPC-interest on three latent variable analyses from the literature where comparisons across groups were of interest (see the digital appendix for `R` code and data[1]). The final section summarizes the findings and discusses some limitations and future work on the use of the EPC-interest for evaluating invariance hypotheses.

[1] Please see the author's Dataverse study `http://hdl.handle.net/1902.1/21816`

2. ASSESSING THE EFFECT OF MISSPECIFIED INVARIANCE RESTRICTIONS ON SEM PARAMETERS OF INTEREST

A structural equation model is any model $\mathbf{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\mu}(\boldsymbol{\theta})$ that imposes a structure on the population covariance matrix $\mathbf{\Sigma}$ and mean vector $\boldsymbol{\mu}$ of observed variables $\boldsymbol{y}$ as a function of a vector $\boldsymbol{\theta}$ of unknown model parameters (Bollen, 1989). A common parameterization of SEM is the LISREL "all-y" model for group g,

$$\boldsymbol{y}_g = \boldsymbol{\nu}_g + \mathbf{\Lambda}_g \boldsymbol{\eta}_g + \boldsymbol{\epsilon}_g, \tag{1}$$

$$\boldsymbol{\eta}_g = \boldsymbol{\alpha}_g + \mathbf{B}_g \boldsymbol{\eta}_g + \boldsymbol{\zeta}_g, \tag{2}$$

where $\boldsymbol{\eta}_g$ is a vector of latent variables, and $\boldsymbol{\epsilon}_g$ and $\boldsymbol{\zeta}_g$ are observed and latent variable residuals. The "measurement model" consists of the first equation involving as parameters the vector of intercepts $\boldsymbol{\nu}_g$, the loading matrix $\mathbf{\Lambda}_g$, and the residual variance matrix $\text{Var}(\boldsymbol{y}_g|\boldsymbol{\eta}_g) = \text{Var}(\boldsymbol{\epsilon}_g) := \mathbf{\Psi}_g$. The second equation is the "structural" part of the model with latent intercepts (or means) $\boldsymbol{\alpha}_g$, latent variable regression coefficients $\boldsymbol{B}_g$, and the (residual) variance matrix $\text{Var}(\boldsymbol{\zeta}_g) := \mathbf{\Phi}_g$ as group-specific parameters. Assuming $\boldsymbol{\eta}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{\zeta}$ are mutually uncorrelated, this model produces as moment structure implications

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Phi}(\mathbf{I} - \mathbf{B})^{-T}\mathbf{\Lambda}' + \mathbf{\Psi}, \tag{3}$$

for the covariances and

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\nu} + \mathbf{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} \tag{4}$$

for the means, where for notational convenience the group-specific parameters have been stacked to obtain the block-diagonal covariance matrix over all groups.

Estimation of the parameters $\mathbf{a} = [\boldsymbol{\nu}', \boldsymbol{\alpha}', (\text{vec}\, \mathbf{\Lambda})', (\text{vech}\, \mathbf{\Phi})', (\text{vec}\, \mathbf{B})', (\text{vech}\, \mathbf{\Psi})']'$

then proceeds by minimizing a nonnegative fitting function $F(\mathbf{S}, \mathbf{m}; \mathbf{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\mu}(\boldsymbol{\theta}))$ so that $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} F(\mathbf{S}, \mathbf{m}; \mathbf{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\mu}(\boldsymbol{\theta}))$ (e.g. Satorra, 1989). A common choice of fitting function is that corresponding to maximum likelihood estimation under the assumption that $\mathbf{y}$ is independently and identically distributed as multivariate normal (Bollen, 1989). Under the model, however, such distributional assumptions do not affect consistency of the parameter estimates (Satorra, 1989).

Models freely estimating all parameters in $\mathbf{a}$ are generally not identifiable. Therefore, restrictions $\mathbf{a} = \mathbf{a}(\boldsymbol{\theta})$ are imposed, such as setting certain loadings to zero, restricting the residual variance matrices to be diagonal, or allowing only recursive latent variable regression coefficients. Any subset of $\mathbf{a}$ may play the role of the parameter interest; attention may focus on differences in the latent means $\boldsymbol{\alpha}_g$ over groups g, for instance, or on differences in latent variable regressions $\mathbf{B}_g$. Although there is in principle no restriction on what may defined as a parameter of interest, often the parameters of direct interest are the structural parameters (Fan and Hancock, 2006).

To identify differences over groups in structural parameters pertaining to latent variables, however, it is necessary to impose cross-group equality restrictions on the measurement parameters: the measurement invariance restrictions. Even when no explicit restrictions are made, but, for instance, one loading is set to unity in each group to identify regression coefficients in each group, the assumption of invariance of such reference loadings is implicit (Hancock, Stapleton and Arnold-Berkovits, 2009). For identifiability of the comparison of interest, then, it is necessary to include in the restrictions $\mathbf{a} = \mathbf{a}_0(\boldsymbol{\theta})$ equality restrictions on the measurement model. In general, equality of loadings $\mathbf{\Lambda}_{g'} = \mathbf{\Lambda}_g \forall g \neq g'$ ("metric invariance") is required for identification of parameters pertaining to the covariance matrix of the latent variables such as $\mathbf{B}_g$, while for identification of latent mean differences in $\boldsymbol{\alpha}_g$, equality of both loadings and intercepts $\mathbf{\Lambda}_{g'} = \mathbf{\Lambda}_g; \boldsymbol{\nu}_{g'} = \boldsymbol{\nu}_g \forall g \neq g'$ ("scalar invariance") is required. Of course, such restrictions may be misspecified, and the misspecifications may cause bias in the parameters of interest (Yuan, Marshall and Bentler, 2003; Millsap, 2007; Kolenikov, 2009). It is not necessary that the full measurement parameter vectors be equal however: partial invariance of at least two indicators per latent concept suffices for identification of the parameters of interest (Byrne, Shavelson and Muthén, 1989). This suggests that after estimation of a fully invariant model with restrictions $\mathbf{a}_0(\boldsymbol{\theta})$, a partially invariant alternative model $\mathbf{a}_a(\boldsymbol{\theta})$ can be considered which frees one equality restriction. Alternatively, $\mathbf{a}_0(\boldsymbol{\theta})$ may itself be a partially invariant model, assuming that it and the alternative model remain identifiable.

The EPC-interest of an equality-constrained measurement parameter with respect to a parameter of interest in the $\mathbf{a}_0$ model is defined as a consistent estimate of the expected change in the parameter of interest if the equality constraint were freed in the $\mathbf{a}_a$ model. Let the parameters of interest $\boldsymbol{\pi}$ be defined as $\boldsymbol{\pi} = \mathbf{P}\theta$. Typically $\mathbf{P}$ is a selection matrix, but $\mathbf{P}$ may also produce a linear combination or contrast of free model parameters, for example the cross-group differences in regression coefficients. Let $\mathbf{A}_a = \partial \mathbf{a}_a / \partial \boldsymbol{\theta}'$. Then $\mathbf{A}_a$ is a logical (0/1) matrix corresponding to the alternative model including the possible misspecification under consideration as though it were a free parameter. Thus $\mathbf{A}_a$ augments the model with an additional parameter – here only freed equality restrictions are examined, but cross-loadings or error covariance can in principle also be incorporated. As shown in the appendix,

$$\text{EPC-interest} := \boldsymbol{\pi} - \hat{\boldsymbol{\pi}} \approx \mathbf{P}(\mathbf{A}_a' \mathbf{J}(\hat{\boldsymbol{\theta}}) \mathbf{A}_a)^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}) \mathbf{A}_a. \tag{5}$$

where $\mathbf{g}(\hat{\boldsymbol{\theta}})$ and $\mathbf{J}(\hat{\boldsymbol{\theta}})$ are consistent estimates of respectively the gradient and the hessian of the fitting function with respect to the unrestricted parameter vector ($\mathbf{a}$), evaluated at the sample parameter estimates under the $\mathbf{a}_0$ model (Satorra, 1989; Bentler and Chou, 1992). For the LISREL all-y model, Neudecker and Satorra (1991) provided $\mathbf{g}$ and $\mathbf{J}$ as a function of the parameter estimates: the EPC-interest depends only on the parameter estimates from the restricted $\mathbf{a}_0$ model.

A key assumption in deriving the EPC-interest is that the hessian $\mathbf{J}$ is approximately constant between the null and alternative model population parameter values (Satorra, 1989). This implies that the alternative model $\mathbf{a}_a$ should not itself be strongly misspecified. Some degree of misspecification in the alternative model is allowed for; in this case the EPC-interest becomes an approximation to the shift in the parameter of interest if the equality restriction were freed. The following section will study the influence of alternative model misspecification on the EPC-interest statistic in an example, and suggests that the approximation may be rather robust to this assumption.

The EPC-interest (equation 5) allows the researcher to fit the invariance model and obtain consistent estimates of the effect of various restrictions on the parameters of interest. Thus, it is similar to the more familiar EPC (Saris, Satorra and Sörbom, 1987) in the sense that it gives an estimated shift in a parameter when freeing a restriction, the difference being that this shift is not in the restricted parameter itself but in the parameter(s) of interest.[2] To avoid confusion with the EPC, we will denote that measure as the "EPC-self" and its standardized version (Kaplan, 1989; Chou and Bentler, 1993) as the "SEPC-self". Since partial invariance testing is done precisely because differences in measurement parameters may affect the parameters of interest, the EPC-interest should prove useful when evaluating whether particular equality restrictions should be maintained or not.

[2] If the parameter of interest were defined to be the restricted parameter itself, EPC-interest will equal EPC-self.

3. ACCURACY OF THE EPC-interest: MONTE CARLO SIMULATION AND POPULATION ROBUSTNESS

A small simulation study evaluates the performance of the EPC-interest in small samples as an estimate of the shift in a parameter of interest when freeing a measurement parameter. In this simulation, a two-group, one-factor model with three indicators was formulated. This model was chosen because it would be just-identified without equality restrictions; therefore equality restrictions are necessarily the only misspecifications. This allows us consider purely violations of scalar or metric invariance and not of other model assumptions. The parameter of interest was taken to be the latent mean difference between the two groups, which was set to 0.2. The unstandardized factor loadings were chosen to equal 1 for all indicators and in both groups, the latent variables were chosen to have variance equal to 1 in both groups, and the error variances of the three indicators were set to 0.5. Thus, in standardized terms the loadings equaled $1/(1+0.5) = 0.667$. Two indicators' intercepts were set to zero in both groups, but the third indicator's intercept violated scalar invariance.

The cross-group difference in the intercept ν_1 (misspecification) was varied across simulation conditions ($\Delta\nu_1 \in \{0.1, 0.3, 0.8\}$). The conditions also varied the number of observations for each of the two groups ($n_g \in \{50, 100, 500\}$). For each of the nine resulting conditions, 200 samples were drawn from multivariate normal distributions based on the population model. In each sample, the misspecified full scalar invariance model was fit to the data. The EPC-self was then calculated for the misspecified parameter, as well as the EPC-interest of the misspecified parameter with respect to the latent mean difference. Table 2 shows the results of this simulation for each of the nine conditions.

Table 2 shows that the bias in the latent mean difference, $\Delta\hat{\alpha}$ (column 5), is affected

Table 2: Monte Carlo simulation results show that the EPC-interest approximates the true latent mean bias well, even in small samples.

		Average over 200 replications				
$\Delta\nu_1$	n_g	EPC-self	$\Delta\hat{\alpha}$	$\Delta\hat{\alpha}$ bias	EPC-interest	EPC-interest bias
0.1	50	0.064	0.240	-0.040	-0.034	0.005
0.3	50	0.213	0.313	-0.113	-0.113	-0.001
0.8	50	0.657	0.505	-0.305	-0.401	-0.096
0.1	100	0.058	0.231	-0.031	-0.031	0.000
0.3	100	0.203	0.323	-0.123	-0.109	0.014
0.8	100	0.619	0.492	-0.292	-0.370	-0.077
0.1	500	0.063	0.233	-0.033	-0.033	0.000
0.3	500	0.208	0.307	-0.107	-0.112	-0.005
0.8	500	0.598	0.501	-0.301	-0.349	-0.048

differently in the different conditions, as represented by combinations of sample size (column 1) and intercept misspecification (column 2). As one would expect, the bias is larger in conditions with larger misspecifications. The EPC-interest statistic (column 6) is meant to estimate this bias (column 5) as a result of the misspecification in the intercept ν_1. The average over repeated samples of EPC-interest is indeed very close to the actual bias in the latent mean difference. This means that the EPC-interest gives a close estimate of the shift in latent mean difference estimate if the misspecified ν_1 parameter were freed. For reference, the usual EPC measure, "EPC-self" is given in column 3. It tends to underestimate the true misspecification in the intercept, as can be seen by comparing columns 1 and 3 in Table 2. The estimate of the bias in latent mean difference given by the EPC-interest is close to the true bias even for the small sample size condition in which each group contains only 50 observations.

The Monte Carlo simulation shows that the finite-sample estimate of EPC-interest is close to the population shift in the parameter of interest, even for small samples. However, the quality of the population approximation is affected by the degree of misspecification in the alternative model. If the alternative model is strongly misspecified, the EPC-interest becomes only an approximation to the actual population shift in the

Table 3: Asymptotic robustness of the EPC-interest to misspecification of the alternative model. Shown is the difference between the EPC-interest and the true bias it is meant to approximate.

Misspecification	Mean	Std. dev.	First quartile	Median	Third quartile
Unif(-0.05, 0.05)	-0.005	0.004	-0.008	-0.005	-0.002
Unif(-0.10, 0.10)	-0.006	0.008	-0.012	-0.006	-0.001
Unif(-0.20, 0.20)	-0.006	0.020	-0.019	-0.007	0.005

parameter of interest.

Fan, Thompson and Wang (1999) suggested to define the degree of misspecification as the power of the chi-square test for rejecting $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta})$. This power can be manipulated by adding a random vector $\boldsymbol{\mu}$ to the population implied covariances $\boldsymbol{\sigma}(\boldsymbol{\theta})$, since the power will then equal $\Pr(\chi^2_d(\lambda) > c_\alpha)$, where c_α is the critical value, the noncentrality parameter $\lambda = \boldsymbol{\mu}'\mathbf{U}\boldsymbol{\mu}$ is a quadratic function of the added vector $\boldsymbol{\mu}$, and $\mathbf{U}$ is given by Satorra (1989, p. 138). Asymptotic robustness to the effect of misspecification in the alternative model can therefore be studied by first calculating the population implied covariance matrix and then adding some amount of random variation $\boldsymbol{\mu}$ to these population covariances to reflect the effect of misspecification. The scalar invariance model is then fitted to this shifted population covariance matrix and mean vector, and the EPC-interest calculated on it. Since the EPC-interest may be more biased by some kinds of misspecifications than by others, the advantage of this method of misspecifying the model, besides keeping the power constant as suggested by Fan, Thompson and Wang (1999), is that the type of alternative model misspecification is not fixed.

Table 3 shows the asymptotic robustness of the EPC-interest approximation to misspecification in the alternative model for increasing degrees of misspecification. Again the two-group, single factor three indicator model was used, with a difference in one of the intercepts of 0.3 and a true latent mean difference of 0.2. The population covariance matrix resulting from this model was then perturbed 200 times with uniform random numbers. The minimum and maximum of the perturbations increased in three conditions from 0.05 to 0.20. The scalar invariance model was fitted to each of the resulting perturbed matrices, and the EPC-interest calculated as well as the "estimate" of the latent mean difference of interest. If the EPC-interest is robust to misspecification of the alternative model, it should be close to the bias in the latent mean difference resulting from fitting the scalar invariance model to the perturbed matrix.

Table 3 shows the mean, standard deviation, and quartiles over the 200 perturbations of the difference between EPC-interest and the bias in the latent mean difference. It can be seen from the increase in standard deviation that this difference can increase with the amount of misspecification in the alternative model. However, the error in the approximation of the EPC-interest is in general small compared to the latent mean difference bias, which ranged between -0.30 and +0.05. This shows that even though the quality of the approximation in principle depends on the closeness of the alternative model to the true model, at least in the example studied this effect is minimal and the EPC-interest appears asymptotically robust to this assumption.

4. EXAMPLE 1: MEAN LEVELS OF DEMOCRACY FACTORS

Armstrong (2011) presented a confirmatory factor analysis of seven indicators of the "level of democracy" obtained from Freedom House. Values for these seven indicators were observed for 193 countries in four subsequent years (2006–2009). Based on substantive decisions made by Freedom House, Armstrong (2011) estimated a model with two separate factors; a maximum-likelihood analysis of this model is shown in Table 4.

The author allowed for differences over time in the loadings and intercepts. If intercepts and loadings are not equal over time, latent mean differences over time would

Table 4: Intercept estimates and standardized loadings from the scalar invariance confirmatory factor analysis. Chi-square: 104, df = 70 ($p = 0.005$), CFI = 0.997, RMSEA = 0.050 ($p = 0.468$), SRMR = 0.010.

Indicator	Intercept	(s.e.)	Loading	(s.e.)
Political rights				
Electoral Process (A)	7.7	(0.303)	4.14	(0.218)
Political Pluralism (B)	10.2	(0.369)	5.06	(0.264)
Functioning of Government (C)	6.6	(0.251)	3.40	(0.182)
Civil liberties				
Freedom of Expression (D)	11.6	(0.307)	4.21	(0.220)
Associational Rights (E)	8.0	(0.267)	3.67	(0.192)
Rule of Law (F)	8.7	(0.320)	4.30	(0.233)
Personal Autonomy (G)	9.8	(0.267)	3.57	(0.196)

Note: Items A, C, and E were measured on a scale from 0–12; items B, D, F, and G ranged from 0–16.

not be identifiable (Steenkamp and Baumgartner, 1998). Therefore invariance testing is performed by estimating a model constraining both intercepts and loadings to be equal over time: the so-called "scalar invariance" model. Measurement invariance testing then consists of comparing the scalar invariance model fit with the fit for a model in which intercepts are allowed to vary over time, but loadings are constrained to be equal ("metric invariance"), and the model in which both loadings and intercepts are freed. None of the chi-square difference tests between these models are statistically significant ($\Delta\chi^2 = 6.21, df = 30$), differences in CFI and RMSEA are below the cutoffs recommended by Chen (2007), and the overall model fit, shown in Table 4, would be judged adequate.

The differences over time in the latent means of Political Rights and Civil Liberties may be of interest. These changes are plotted in Figure 1 based on the scalar invariance model. In Figure 1, the year 2006 is taken as the reference group by setting its estimate to zero. Figure 1 shows that, assuming measurement invariance over time, no changes in either of the factors are observed in this period. The EPC-interest can now be applied to assess the sensitivity of these changes to the scalar invariance assumption.

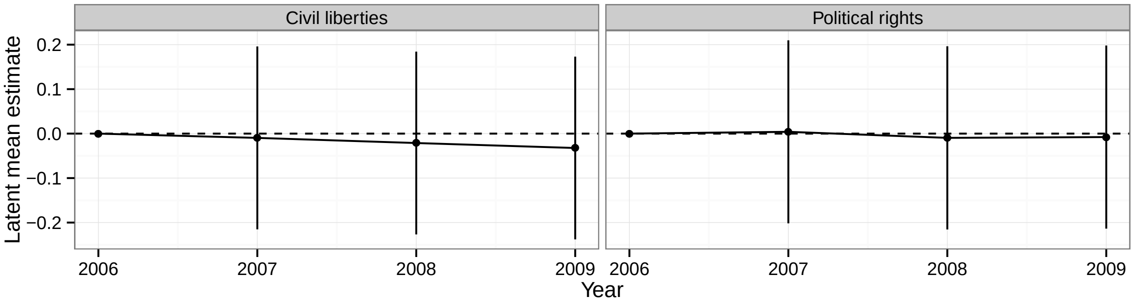

Figure 1: Change in "level of democracy" over time based on scalar invariance model of Freedom House data from 193 countries using the year as grouping variable. Latent means of factors Civil Liberties and Political Rights with two standard error-intervals. The dotted reference line indicates no average change relative to 2006.

All the |EPC-interest| < 0.008. This means that none of the estimates of change in the latent mean differences over time would change by more than 0.008 in absolute value if a scalar invariance restriction on loadings or intercepts were freed. The EPC-interest therefore yields much the same conclusion as the measurement invariance tests. It also provides a reasoning behind selecting the scalar invariance model: under the scalar invariance model, the parameters plotted in Figure 1 are identifiable and any effects of misspecification in the scalar invariance restrictions is too slight to substantially change Figure 1.

Another analysis of interest to Armstrong (2011) was the comparison of democracy levels across countries with different levels of press freedom (low, middle, high). In this case the relevant grouping variable is not the year, but the press freedom variable. The same logic applies to this comparison as to the over-time comparison: measurement parameters (intercepts and loadings) should be the same for countries with low, middle, and high amounts of press freedom if these groups are to be comparable. Contrary to the across-time invariance test, however, imposing scalar invariance restrictions leads to a significantly worse model fit in terms of chi-square ($\Delta\chi^2 = 747, df = 21$). The CFI's for the free, metric invariance, and scalar invariance models are 0.900, 0.848, and 0.682

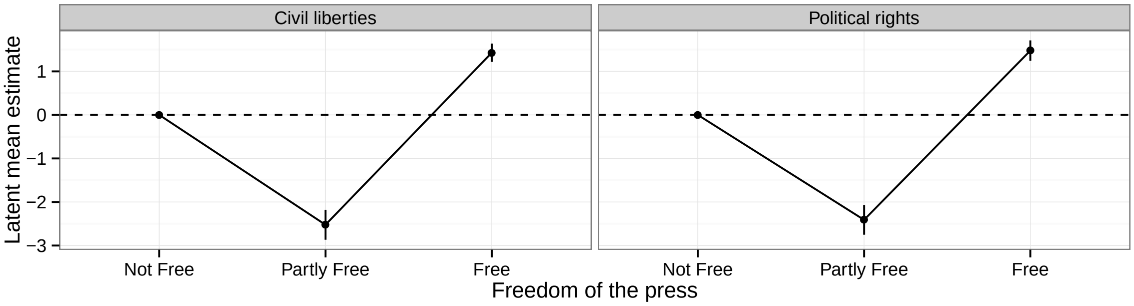

Figure 2: Latent means of Civil Liberties and Political Rights factors by level of freedom of the press based on the scalar invariance model. The dotted reference line indicates no difference relative to the reference “not free” group.

respectively, while the corresponding RMSEA’s are 0.232, 0.251, 0.322 respectively. This indicates that the scalar and metric invariance restrictions fit badly, suggesting that, strictly speaking, the regression of the Civil Liberties and Political Rights scores on Freedom of the Press is not valid, since the comparison over press freedom groups is possibly confounded with measurement differences. Figure 2 plots the latent mean estimates ignoring the lack of model fit of the scalar invariance model which allows for identification of these differences.

Invariance testing has indicated that Figure 2 may not provide a valid comparison, because it is based on a misspecified measurement invariance model. The EPC-interest allows us to investigate whether these misspecifications are capable of changing the substantive conclusion of interest shown in Figure 2, namely that there appears to be a nonlinear relationship between freedom of the press and the level of democracy. Figure 2 shows that an EPC-interest in the “partly free” group of at least 2, and an EPC-interest in the “free” group of at least 1 in absolute value would be required to change the substantive conclusion.[3] Table 5 shows both columns of EPC-interest values that involve a EPC-interest of at least 1 in absolute value. Although the EPC-interest

[3] Note that this requirement is very similar to the bounds derived by Imai and Yamamoto (2010).

Table 5: EPC-interest of equality-constrained parameters with respect to latent mean differences shown in Figure 2.

		EPC-interest when freeing...	
Mean estimate of...	Group	`F~1` in group “Free”	`D~1` in group “Partly free”
Political rights	Free	-0.247	0.064
Civil liberties	Free	0.021	0.095
Political rights	Partly free	-1.125	-1.037
Civil liberties	Partly free	-0.681	-0.561

values are much larger than those found for the scalar invariance model with respect to time, none of the EPC-interest values are large enough to change the substantive conclusions of interest. In spite of the obvious model misspecification, therefore, the comparison between groups representing different levels of press freedom does not appear to be threatened by differences in measurement parameters.

5. EXAMPLE 2: REGRESSION COEFFICIENTS IN 19 COUNTRIES

While the comparison of latent means requires scalar invariance, the cross-group comparison of regression coefficients among latent variables calls for metric invariance (e.g. Steenkamp and Baumgartner, 1998). A study by Davidov et al. (2008) on the effect of human values on attitudes toward immigration illustrates the application of the EPC-interest to this more complex situation. The authors compare 19 European countries on four regression coefficients between latent variables measured by 17 items. The effect of $19 \times 17 = 323$ partial metric invariance restrictions' potential effects on $19 \times 4 = 76$ country-specific regression coefficients needs to be assessed: $323 \times 76 = 24,548$ EPC-interest statistics. From this apparently daunting task, the EPC-interest applied to this example produces a surprisingly simple picture: it turns out to be inconsequentially small in all but four cases.

Figure 3 reproduces Davidov et al. (2008)'s model along with the metric invariance model's cross-country average structural regression coefficients. Interest focuses on the structural regression coefficients representing the effect of the two value dimensions "Self-transcendence" and "Conservation" advanced by Schwartz and Bilsky (1987) on two different dimensions of attitudes toward immigration, "Allow" and "No conditions". In this parametrization of the metric invariance model, the first country's latent variables have been standardized and the other countries' latent variable variances allowed to vary to take heteroskedasticity into account. This parametrization is equivalent to the more common practice of fixing one loading per latent variable to unity in each country and facilitates the interpretation and comparison of the regression coefficients over countries.

Data were obtained from the European Social Survey 2002, a high quality cross-national probability survey (Jowell et al., 2007). Each latent variable is measured with multiple observed indicators. Davidov et al. (2008) compared 19 countries on these regression coefficients: Austria (n =2,257), Belgium (1,899), Czech Republic (1,360), Denmark (1,506), Finland (2,000), France (1,503), Germany (2,919), Great Britain (2,052), Greece (2,566), Hungary (1,685), Ireland (2,046), Netherlands (2,364), Norway (2,036), Poland (2,110), Portugal (1,510), Slovenia (1,519), Spain (1,729), Sweden (1,999), and Switzerland (2,037). For the original data, precise wording of the questions, and further information on data collection procedures, we refer to the ESS website.[4] The left-hand side of Table 6 presents descriptive statistics for the 17 observed variables. The within-country and between-country standard deviations call attention to the considerable between-country variation in the means, which could reflect substantive differences between the 19 countries, but could also originate in cross-country loading (or intercept) differences.

Following standard practice, the original authors fit the full metric invariance model

[4] http://ess.nsd.uib.no/ess/round1/

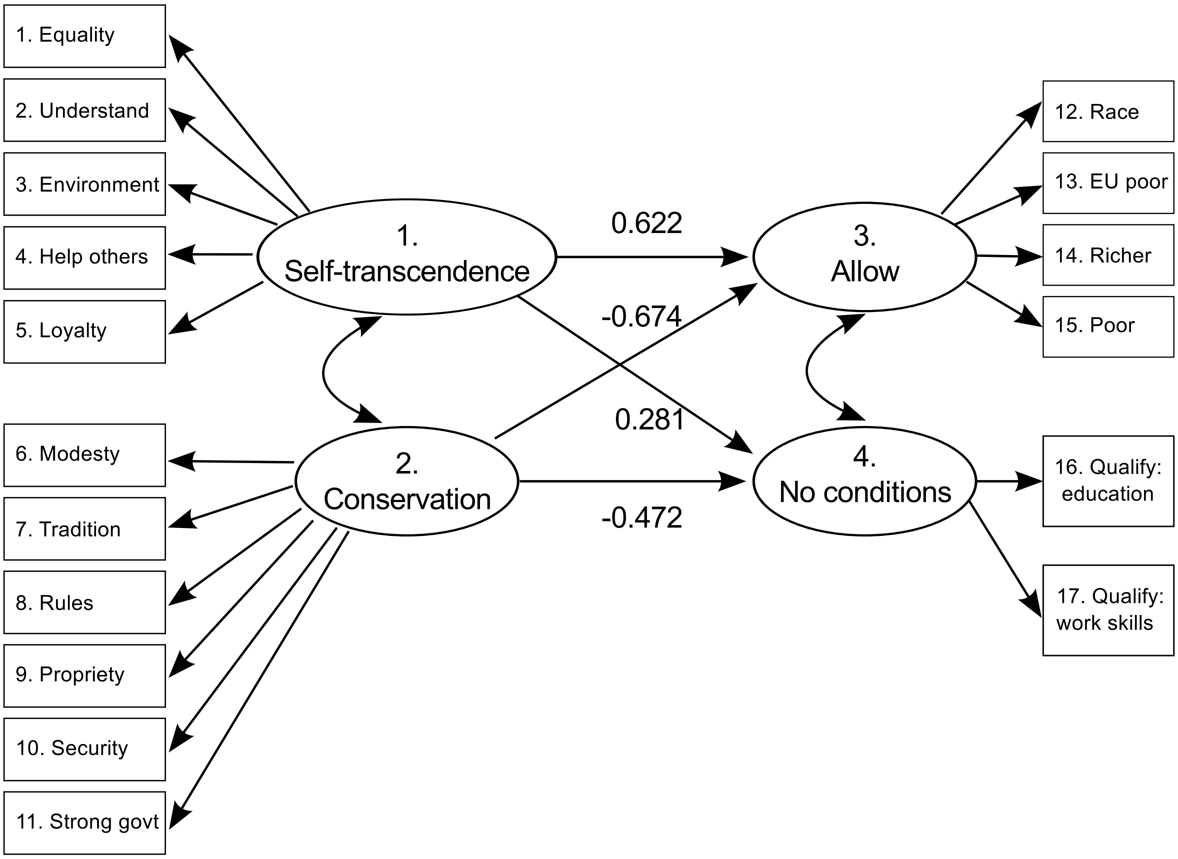

Figure 3: Structural equation model of relationships of interest between four latent variables. The regression coefficient estimates shown are averages over all 19 countries under the full metric invariance model (see Figure 4 for country estimates).

Table 6: Left: descriptive statistics for the observed variables. Overall means over countries, within-country standard deviation, and between-country standard deviation. Right: loading estimates in the full metric invariance model.

		Overall	Within-	Between-	Self-	Conserv.	Allow	No
		mean	country	country	trans	2.	3.	cond.
			sd.	sd.	1.			4.
1.	Equality	2.07	(1.04)	(0.18)	0.648			
2.	Understanding	2.41	(1.05)	(0.18)	0.711			
3.	Environment	2.13	(0.99)	(0.19)	0.697			
4.	Helping others	2.31	(0.99)	(0.16)	0.734			
5.	Loyal to friends	1.97	(0.89)	(0.19)	0.647			
6.	Modesty	2.83	(1.24)	(0.42)		0.677		
7.	Tradition	2.76	(1.32)	(0.35)		0.768		
8.	Follow rules	3.14	(1.34)	(0.37)		0.847		
9.	Proper behavior	2.70	(1.23)	(0.31)		0.920		
10.	Security	2.37	(1.18)	(0.36)		0.801		
11.	Govt. strong	2.43	(1.19)	(0.37)		0.838		
12.	Allow other race	2.54	(0.78)	(0.27)			0.589	
13.	Allow EU poorer	2.45	(0.76)	(0.28)			0.612	
14.	Allow richer	2.48	(0.81)	(0.21)			0.517	
15.	Allow poorer	2.51	(0.77)	(0.28)			0.632	
16.	Qualify education	6.22	(2.64)	(0.64)				1.732
17.	Qualify work skills	6.75	(2.65)	(0.78)				2.170

to the ESS data: a multiple group structural equation model in which the loadings in Figure 3 are constrained to be equal across the 19 countries, while the structural regression coefficients and other parameters are allowed to vary. The right-hand side of Table 6 gives the resulting estimates of the loadings (under the variance parameterization). Parameter estimates of interest for the 19 countries are shown in Figure 4.

Davidov et al. (2008, 589) test for (partial) metric equivalence by examining overall fit measures as well as MI and SEPC-self. They decide that "the overall fit measures (CFI = 0.95, NFI = 0.94, RMSEA = 0.01, Pclose = 1.0) suggest that [the full metric invariance model] is acceptable. However, modification indices pointed to misspecifications in the model. Therefore, in model 2 we (...) had to relax the measurement invariance constraints for some items." Based on the MI and (S)EPC-self, the author's final model frees four out of the possible 323 loading equalities in three different countries, namely "Conservation → Modesty" in Portugal, "Conservation → Govt. strong" in Ireland, and "Self-transcendence → Loyal to friends" in Portugal and Denmark.[5]

The EPC-interest statistic provides an alternative way to assess whether a loading equality restriction is substantively important or not. Saris, Satorra and Van der Veld (2009) suggested a cutoff of 0.1 in absolute value for correlations and standardized regresssion coefficients. Given the parameterization used, this appears to be a reasonable criterion for the EPC-interest with respect to regression coefficients as well. Although there are potentially 76 affected parameters of interest for each of the 323 equality restrictions, as it happens only the four EPC-interest statistics shown in Table 7 meet this criterion. The form of the model plays an important role here: equality constraints on loadings in one country hardly affect structural parameters in another, misspecified constraints on one dependent latent variables' loadings hardly affect the regression coefficients of the other, and latent variables with many indicators are generally less

[5] Davidov (2012, personal communication).

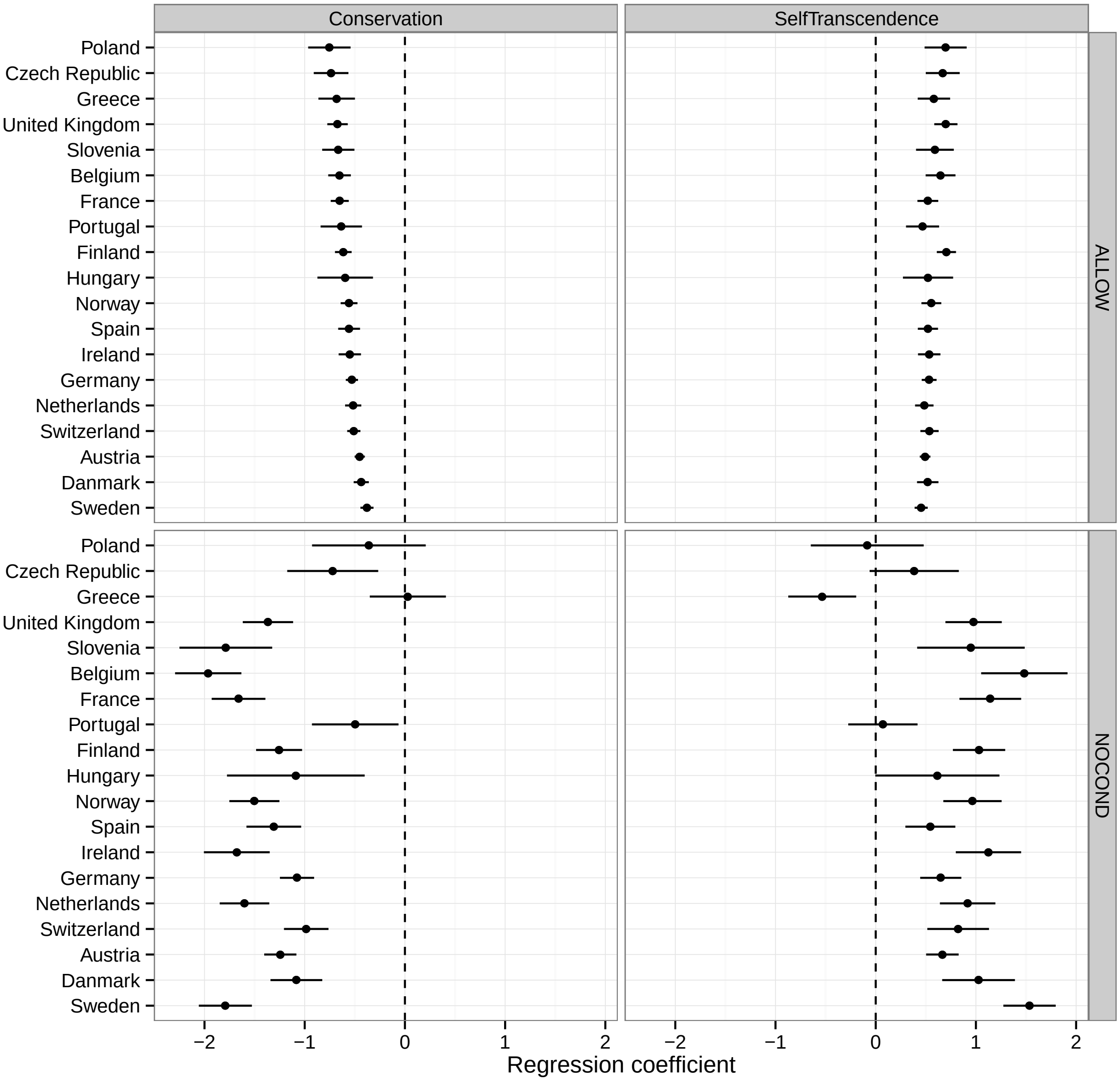

Figure 4: Estimates ±2 s.e. for the four regression coefficients between latent variables of interest under the metric invariance model in 19 countries.

Table 7: EPC-interest statistics of at least 0.1 in absolute value with respect to the latent variable regression coefficients. For reference, the standardized expected parameter change (SEPC-self) is also given.

	Loading "Conditions → Qualify work skills" in...			
	Slovenia	France	Hungary	Ireland
EPC-*interest with respect to:*				
Conditions → Self-transcendence	-0.073	-0.092	-0.067	0.073
Conditions → Conservation	0.144	0.139	0.123	-0.113
SEPC-self	0.610	0.692	0.759	-0.514

affected by misspecifications.

In summary, even though the full metric invariance model shows "close fit" in terms of CFI and RMSEA, this model still contains misspecified equality restrictions on loadings that threaten the comparison of the regression coefficients of interest. Freeing these loadings takes this possibility into account while still allowing for the identification of the parameters of interest. The misspecified restrictions on loadings in question were not detected with the SEPC-self and MI: these measures detected other misspecifications that were large in the sense of having strongly differing loadings.

6. EXAMPLE 3: SEX DIFFERENCES IN VALUE PRIORITIES

Men and women differ in their latent value priorities according to a study by Schwartz and Rubel (2005), who offered substantive explanations for this finding. Using the European Social Survey 2002 data introduced above, all 21 observed ESS value indicators (see appendix) were modeled as measuring eight factors, following a scheme detailed by Schwartz and Bilsky (1987). These "universal human values" were: benevolence (BE), universalism (UN), self-direction (SD), stimulation (ST), hedonism (HE), achievement/power (ACPO), security (SE), and conformity/tradition (COTR).

Before comparing men and women on their latent mean value priorities, the threat of sex differences in response behavior should be ruled out. To this end, a full scalar invariance model was judged to be satisfactory by the authors (p. 1013):

> The fit indices for configural and metric invariance were satisfactory (comparative fit index [CFI] = .90, adjusted goodness-of-fit index [AGFI] = .94, root-mean-square residual [RMR] = .07, root-mean- square error of approximation [RMSEA] = .037, confidence interval [CI] = .037.038, p of close fit [PCLOSE] = 1.0). (...) When we constrained scalar invariance, chi-square deteriorated significantly, $\Delta\chi^2(19) = 3313$, $p = .001$, but CFI did not change. Change in chi-square is highly sensitive with large sample sizes and complex models. The other indices suggested that scalar invariance might be accepted (CFI = .88, RMSEA = .04, CI = .039.040, PCLOSE = 1.0).

I re-analyzed these data, fitting the full scalar invariance model to obtain estimates of the differences between the group of 18,519 men and the group of 16,740 women in their eight latent values factor means. Following the original authors, these sex difference estimates were controlled for age, education level, and country.

Figure 5 shows estimates based on the full scalar invariance model as red dots. These may be interpreted as latent effect sizes. Positive values indicate that men value this factor more, while negative values below the dashed line mean women are estimated to value a factor more. It can be seen that Achievement/Power (ACPO), Stimulation (ST), Self-direction (SD), and Hedonism (HE) are estimated to be more important to men, while Security (SE), Universalism (UN), and Benevolence (BE) appear more valued by women. Evolutionary psychology was suggested by Schwartz and Rubel (2005) as an explanation for these differences. For instance: "stimulation values emphasize and justify the pursuit of excitement, novelty, and challenge in life. Men engage in risky behavior more than women do (...) From an evolutionary perspective, they do this

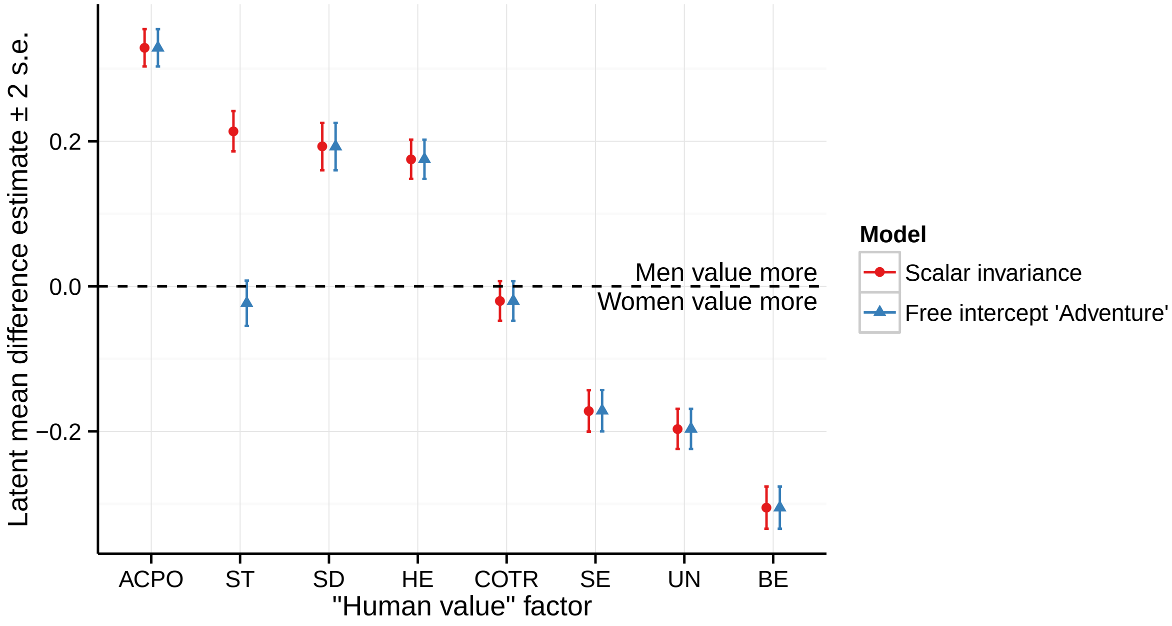

Figure 5: Sex differences in value priorities before and after allowing for between-sex differences in the intercept of “Important to seek adventures and have an exiting life”.

because it serves to increase their social status” (p. 1021).

It may appear that such conclusions on sex differences in human values are protected from the threat of the alternative explanation that men and women answer the questions differently by the “closeness” of the full scalar invariance model fit as evidenced by CFI, RMSEA, etc. But “close fit” does not guarantee invulnerable substantive conclusions. Sex difference estimates may change when misspecified intercept equality restrictions are freed–the question is how much. The EPC-interest for the change in Stimulation’s latent mean difference after freeing the intercept equality restriction on the item (*Adventure*) “important to seek adventures and have an exiting life” is -0.243, for example. Since the effect size estimate in Figure 5 of this sex difference is +0.214 (s.e. 0.0139) under the full scalar invariance model, the EPC-interest indicates that freeing this intercept equality has the potential of reversing the conclusion: women would be estimated to value Stimulation more under the alternative model that frees this misspecification, according to the EPC-interest.

As detected by the EPC-interest, freeing this intercept equality in an alternative model does indeed reverse the conclusion on Stimulation: estimates of the sex differences in latent means after fitting this alternative model are shown as blue triangles in Figure 5. The improvement in model chi-square relative to full scalar invariance (likelihood ratio test) is $\chi^2_1 = 827$ ($p < 10^{-6}$). It is immediately apparent from Figure 5 that freeing the intercept of item *Adventure*, which measures Stimulation, changes the estimate of sex differences in that factor. For Stimulation, although the effect size estimate of -0.021 (s.e. 0.014) does not differ significantly from zero, conclusions would now be reversed, women valuing Stimulation more than men: a different evolutionary explanation may have to be found.

In short, the original authors, encouraged by "close fit", assumed scalar invariance, concluded that men value Stimulation more than women and theorized that this was due to natural selection of men who take risks. EPC-interest estimates for the scalar invariance model warn that different assumptions with the same data permit the researcher to conclude the opposite.

7. DISCUSSION AND CONCLUSION

Invariance of measurement parameters is a requirement for cross-group comparisons of structural SEM parameters with the purpose of ruling out confounding of measurement differences with structural differences. However, as shown in the second column of Table 1, when the invariance model does not fit the data, this need not necessarily invalidate the comparison of interest. The first two example analyses, which unnecessarily discarded data or freed parameters, attested to this point. Moreover, as also shown in Table 1, the invariance model may appear to fit well – leading the researcher to ignore certain “small” misspecifications – even though disregarded misspecifications can still have a substantial impact on the comparison of interest. In the third example, the differences between men and women in their valuation of “Stimulation” was estimated as being either positive or negative, depending on whether an intercept was freed over groups or not. That intercept’s equality had appeared to fit “closely” in earlier analyses, however, possibly due to the large sample size.

In these examples, and in a small simulation study, the EPC-interest statistic was found to provide a reasonable measure of whether misspecifications in invariance constraints are substantively relevant. It approximates the change in parameters of interest if equality-constrained measurement parameters were freed. `R` code (R Core Team, 2012) implementing the EPC-interest for the SEM library `lavaan` (Rosseel, 2012) and allowing for reproduction of the examples discussed in this paper is available online at `http://hdl.handle.net/1902.1/21816`.

Although I am confident in the demonstrations given here, the simulation study was not meant to provide a full evaluation under a wide range of realistic conditions. Such an evaluation may provide more insight in the application of the EPC-interest. Similarly, a systematic comparison of its practical application with that of other invariance testing methods for a large range of topics, measurement models, structural models, misspecifications, and parameters of interest was necessarily outside of the scope of this paper.

A disadvantage of the univariate EPC-interest measure used is that the measurement parameters in latent variable models are often dependent, meaning that the EPC-interest suffers from similar problems of stepwise model improvement and possible capitalization on chance as the EPC-self and MI; MacCallum, Roznowski and Necowitz (1992) discussed these problems, noting that modifications in step-wise improvement made without the aid of theory often do not replicate across samples, and that cross-validation could not provide a solution, although more positive results when using MI and EPC were found by Kwok, Luo and West (2010), King (2011, ch. 7), and Whittaker (2012), while Hancock (1999) suggested a method of controlling type-I error that may apply to the test of no change. Furthermore, Hancock, Stapleton and Arnold-Berkovits (2009) showed that the question of "which" loading or intercept in a multiple group SEM is invariant is often not identifiable from the data, but that this need not matter for the comparison of substantive interest. Although these problems are not specific to the EPC-interest, they remain a concern. One topic for future study is therefore the degree to which the stability problems of stepwise model improvement extend to the EPC-interest.

When stability of the schoice of invariant measurement parameters is evaluated, it should likewise be evaluated whether the final comparison made differs across samples. That is, the extent to which situations (2) and (3) of Table 1 are indeed prevented. After all, the goal of the EPC-interest is not to detect precisely *which* measurement parameters are non-invariant, but rather if and how a comparison of structural parameters across groups is warranted. It is not clear whether this goal would be as unstable across samples as the decision on invariance of particular measurement parameters.

This article only considered application of the EPC-interest to scalar and metric invariance, that is, to equality constraints across groups. It was assumed that the basic model held at least approximately, that is, that configural invariance held. Structural parameters are, however, likely to be even more sensitive to configural invariance than to equality restrictions. For example, Yuan, Marshall and Bentler (2003, Table 1) showed that large biases in structural variance parameters can occur due to relatively small misspecification in correlated errors. Cross-loadings are likely to strongly affect structural regression coefficients. The EPC-interest has been defined in such a way as to include the possibility of freeing other restrictions than scalar and metric invariance by the choice of $\mathbf{A}_a$ in Equation 5. An evaluation of the use of the EPC-interest in this case would be a useful topic for future study, effectively extending the evaluation made in this paper from scalar and metric invariance to configural invariance.

In the sensitivity analysis approach, partial measurement invariance is no longer a property of a measurement model, but becomes a property holding only with respect to a particular analysis of interest. Use of the EPC-interest for measurement invariance evaluation requires more substantive knowledge and input from the researcher. Whether this is a drawback or an advantage is up for debate: arguably, invariance is a requirement in comparative research because of its substantive implications, so substantive implications should guide decisions on whether or not equality restrictions are warranted. As demonstrated by the example analyses, invariance tests do not guarantee that "closely fitting" but misspecified restrictions are inconsequential for parameters of interest, nor that "badly fitting" equality restrictions are substantively relevant. The EPC-interest thus represents a trade-off between higher analysis complexity and a greater degree of confidence that the groups are comparable for the purposes at hand.

REFERENCES

Armstrong, D.A. 2011. "Stability and change in the Freedom House political rights and civil liberties measures." *Journal of Peace Research* 48:653–662.

Bartholomew, D.J., M. Knott and I. Moustaki. 2011. *Latent variable models and factor analysis: a unified approach.* New York: John Wiley & Sons.

Bentler, PM and C.P. Chou. 1992. "Some new covariance structure model improvement statistics." *Sociological Methods & Research* 21:259–282.

Bollen, K.A. 1989. *Structural Equations with Latent Variables.* New York: John Wiley & Sons.

Byrne, B.M., R.J. Shavelson and Bengt Muthén. 1989. "Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance." *Psychological Bulletin* 105:456.

Chen, F.F. 2007. "Sensitivity of goodness of fit indexes to lack of measurement invariance." *Structural Equation Modeling* 14:464–504.

Cheung, G.W. and R.B. Rensvold. 2002. "Evaluating goodness-of-fit indexes for testing measurement invariance." *Structural Equation Modeling* 9:233–255.

Chou, C.P. and PM Bentler. 1993. "Invariant standardized estimated parameter change for model modification in covariance structure analysis." *Multivariate Behavioral Research* 28:97–110.

Clinton, J., S. Jackman and D. Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98:355–370.

Davidov, E. 2009. "Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective." *Political Analysis* 17:64–82.

Davidov, E., B. Meuleman, J. Billiet and P. Schmidt. 2008. "Values and support for immigration: a cross-country comparison." *European Sociological Review* 24:583–599.

Fan, Weihua and Gregory R Hancock. 2006. "Impact of post hoc measurement model overspecification on structural parameter integrity." *Educational and psychological measurement* 66:748–764.

Fan, Xitao, Bruce Thompson and Lin Wang. 1999. "Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes." *Structural Equation Modeling: A Multidisciplinary Journal* 6:56–83.

French, B.F. and W.H. Finch. 2006. "Confirmatory factor analytic procedures for the determination of measurement invariance." *Structural Equation Modeling* 13:378–402.

Hancock, Gregory R. 1999. "A sequential Scheffé-type respecification procedure for controlling type I error in exploratory structural equation model modification." *Structural Equation Modeling: A Multidisciplinary Journal* 6:158–168.

Hancock, Gregory R., Laura M. Stapleton and Ilona Arnold-Berkovits. 2009. "The tenuousness of invariance tests within multisample covariance and mean structure models." In *Structural Equation Modeling in Educational Research: Concepts and Applications*, ed. T. Teo and M.S. Khine. Rotterdam, The Netherlands: Sense Publishers pp. 137–174.

Hausman, J.A. 1978. "Specification tests in econometrics." *Econometrica: Journal of the Econometric Society*.

Hu, L. and P.M. Bentler. 1998. "Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification." *Psychological methods* 3:424.

Imai, K. and T. Yamamoto. 2010. "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science* 54:543–560.

Jackman, S. 2001. "Multidimensional analysis of roll call data via Bayesian simulation:

identification, estimation, inference, and model checking." *Political Analysis* 9:227–241.

Jowell, Roger, Caroline Roberts, Rory Fitzgerald and Gillian Eva. 2007. *Measuring Attitudes Cross-nationally: Lessons from the European Social Survey.* Thousand Oaks, CA: SAGE.

Kaplan, D. 1989. "Model modification in covariance structure analysis: Application of the expected parameter change statistic." *Multivariate Behavioral Research* 24:285–305.

King, BL. 2011. *Unbiased measurement of health-related quality-of-life.* Unpublished PhD dissertation, University of Amsterdam.
URL: *http://dare.uva.nl/document/342053*

Kolenikov, S. 2009. "Biases of parameter estimates in misspecified structural equation models." *Sociological Methodology* 41:119–157.

Kwok, Oi-Man, Wen Luo and Stephen G West. 2010. "Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study." *Structural Equation Modeling* 17:216–240.

Lord, F. M. and M. R. Novick. 1968. *Statistical theories of mental scores.* Reading: Addison–Wesley.

MacCallum, R.C., M. Roznowski and L.B. Necowitz. 1992. "Model modifications in covariance structure analysis: the problem of capitalization on chance." *Psychological Bulletin; Psychological Bulletin* 111:490.

Magnus, J.R. and A.L. Vasnev. 2007. "Local sensitivity and diagnostic tests." *The Econometrics Journal* 10:166–192.

Meredith, W. 1993. “Measurement invariance, factor analysis and factorial invariance.” *Psychometrika* 58:525–543.

Meuleman, B. 2012. “When are item intercept differences substantively relevant in measurement invariance testing?” In *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*, ed. S. Salzborn, E. Davidov and J. Reinecke. Wiesbaden, Germany: Springer-Verlag pp. 97–104.

Millsap, R.E. 1997. “Invariance in measurement and prediction: Their relationship in the single-factor case.” *Psychological Methods* 2:248.

Millsap, R.E. 2007. “Invariance in measurement and prediction revisited.” *Psychometrika* 72:461–473.

Millsap, R.E. and H.T. Everson. 1993. “Methodology review: Statistical approaches for assessing measurement bias.” *Applied Psychological Measurement* 17:297–334.

Millsap, R.E. and O.M. Kwok. 2004. “Evaluating the impact of partial factorial invariance on selection in two populations.” *Psychological Methods* 9:93.

Neudecker, H. and A. Satorra. 1991. “Linear Structural Relations: Gradient and Hessian of the Fitting Function.” *Statistics and Probability Letters* 11:57–61.

Oberski, D.L. 2012. “Comparability of Survey Measurements.” In *Handbook of Survey Methodology for the Social Sciences*, ed. Lior Gideon. New York: Springer-Verlag pp. 477–498.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Rosseel, Y. 2012. "`lavaan`: An `R` Package for Structural Equation Modeling." *Journal of Statistical Software* 48:1–36.

Saris, W.E., A. Satorra and D. Sörbom. 1987. "The Detection and Correction of Specification Errors in Structural Equation Models." *Sociological Methodology* 17:105–129.

Saris, W.E., A. Satorra and W.M. Van der Veld. 2009. "Testing structural equation models or detection of misspecifications?" *Structural Equation Modeling* 16:561–582.

Satorra, A. 1989. "Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach." *Psychometrika* 54:131–151.

Schmitt, N. and G. Kuljanin. 2008. "Measurement invariance: Review of practice and implications." *Human Resource Management Review* 18:210–222.

Schwartz, S.H. and T. Rubel. 2005. "Sex differences in value priorities: Cross-cultural and multimethod studies." *Journal of Personality and Social Psychology* 89:1010–1028.

Schwartz, S.H. and W. Bilsky. 1987. "Toward a universal psychological structure of human values." *Journal of Personality and Social Psychology* 53:550.

Steenkamp, JBEM and H. Baumgartner. 1998. "Assessing measurement invariance in cross-national consumer research." *Journal of Consumer Research* 25:78–107.

Treier, S. and S. Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52:201–217.

Vandenberg, R.J. and C.E. Lance. 2000. "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research." *Organizational Research Methods* 3:4–70.

Whittaker, T.A. 2012. "Using the Modification Index and Standardized Expected Parameter Change for Model Modification." *The Journal of Experimental Education* 80:26–44.

Yoon, M. and R.E. Millsap. 2007. "Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study." *Structural Equation Modeling* 14:435–463.

Yuan, K.H., L.L. Marshall and P.M. Bentler. 2003. "Assessing the effect of model misspecifications on parameter estimates in structural equation models." *Sociological Methodology* 33:241–265.