Sensitivity Analysis

Daniel L. Oberski

Department of Methodology & Statistics, Utrecht University, The Netherlands

The goal of cross-cultural analysis

Cross-cultural analysis is all about comparing groups. Whether the groups being compared are countries, ethnicities, religions, or scientific fields, the goal is always to assess whether these groups differ in some respect and to interpret those differences in substantive terms. For example, political scientists may be interested in cross-country differences in "internal political efficacy", the public's feeling that they are personally capable of influencing politics (Coleman & Davis, 1976). Healthy democracies are thought to need high levels of this feeling (Wright, 1981), and it may be of interest to examine why some countries have more of it than others; Karp & Banducci (2008), for instance, argued that proportional versus majoritarian election systems explain some of these found cross-country differences. Other comparisons of substantive interest might regard the *link* between efficacy and voting: what differences are there across countries in the strength of this link and how can we explain those differences?

However, before such substantive conclusions can be drawn, a range of "threats" to their validity (Shadish, Cook, & Campbell, 2002) must be ruled out as alternative explanations. This chapter, and the entire field of "measurement invariance", regards one of those threats: the alternative explanation that seemingly substantive cross-group differences are actually due to differences in measurement. Such differences can occur because in the social sciences perfect measures are rarely available. For example, if Korean respondents tend to give more modest answers than Danish ones, Kim, Helgesen, & Ahn's (2002) substantive finding that Koreans have lower political efficacy than Danes could potentially be "explained away" by such response effects, rather than the actual efficacy differences in which these authors were interested. Therefore, this alternative explanation of Kim et al.'s substantive finding should be ruled out.

The simplest way to rule out this alternative explanation that lack of "measurement invariance" causes the substantive differences is to know the measurement relation between the observed indicator and the true variable of substantive interest. Figure 1 illustrates why. This figure is adapted from a study by Coromina, Saris, & Oberski (2008) on comparing measures of "political interest" across countries. One of the measures of "political interest" was thought to be spending time watching political programs on TV (expressed as a proportion of the total time spent watching TV). Figure 1 plots the estimated relationship between this indicator on the vertical axis and the true variable of substantive interest, "political interest", on the horizontal axis. One solid line corresponds to the estimated relationship in Italy, while the other solid line is the same relationship in all 20 other European countries in the study.

Figure 1 shows that the measurement relations differ strongly across these groups. For example, if we found both groups spent the same proportion of their TV-watching time, say 0.6 (or 60%), watching programs about political issues, we would likely conclude that Italy and the other countries were similar. But looking at the horizontal dashed gray line in Figure 1 reveals that a response of 0.6 is indicative of a much higher interest in political issues in Italy than in the other countries. So an apparent substantive similarity actually corresponds to a considerable true difference. Conversely, the vertical dashed gray line shows that if we found an apparently strong substantive difference in watching political TV programs (the vertical axis) of 0.6 in Italy versus 0.8 in other countries, this would actually correspond to the exact same level of true interest in both countries (0.8 on the horizontal axis).

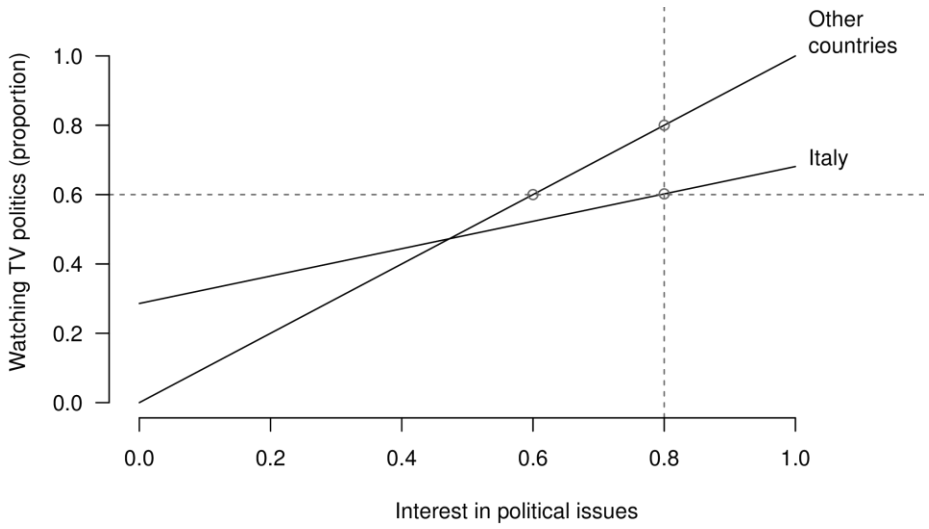

Figure 1. Example of (estimated) non-invariance, adapted from Coromina et al. (2008). Substantive comparisons on the "watching TV politics" indicator do not necessarily translate into similar differences on the true "interest in political issues" variable.

Why measurement invariance testing is not enough

We have seen that if we want to do cross-cultural research by comparing *observed values* across groups, we would need to know their measurement relationships with the *true values*.

Traditional approaches to "measurement invariance" estimate these relationships using factor analysis or IRT models (categorical-data factor analysis) (Steenkamp & Baumgartner, 1998; Schmitt & Kuljanin, 2008; Millsap, 2012). The common logic has been the following: if measurement relationships such as those in Figure 1, i.e. the intercepts and slopes (loadings), are equal across the groups, then the researcher can rest easy with regard to this "threat to validity". Of course, other potential "threats to validity", such as the validity of the concept itself, appropriateness of the factor model, or the external validity of the study as a whole, remain unaffected by such assessments.

In addition, even assuming the factor model holds, the measurement relationships are not typically known exactly but are estimated from a sample. Therefore, a common practice has been to test the null hypothesis that measurement relationships such as the lines in Figure 1 are equal across groups in the unobserved population, by testing whether their parameters (intercepts and slopes) are. If this hypothesis is *not* rejected using a chi-square test, and there was sufficient power to (Saris & Satorra, 1993; Saris, Satorra, & Van der Veld, 2009), classical researchers concluded that the lines can be assumed to equal one another in the population and cross-cultural differences interpreted substantively.

A serious problem with this approach was soon noted by applied researchers in this field. In practice, the null hypothesis that measurement relationships are identical across countries, cultures, genders, etc., was almost always rejected. But just because measurement differences exist does not mean that they seriously disturb the substantive interpretations: they must also be large enough to do so. In short, testing the null hypothesis of measurement invariance fails because there will usually be some differences, however small, in measurement relationships over the groups to be compared. The question then becomes whether these differences are large enough to change the substantive conclusions.

Several methods have been suggested to account for the fact that the null hypothesis of exactly equivalent measurement rarely holds in practice. In particular, the use of "fit measures", such as the Comparative Fit Index (CFI) and Root Mean Squared Error of Approximation (RMSEA) to compare models has been advocated to compare models with and without equality restrictions (Cheung & Rensvold, 2002; Chen, 2007). Such measures do not provide a hypothesis test, but recognize that the model with invariance restrictions may be wrong to some extent. The measures then indicate how large this "wrongness", the model misfit, is, compared to some baseline. For the CFI, this baseline is the "worst possible" model fit (usually the model that states all variables are uncorrelated); for the RMSEA, it is the complexity of the model as measured by its degrees of freedom. See the rest of this volume for more detailed descriptions of these and other approaches to measurement invariance for cross-cultural research. Overall, the idea of these measures in measurement invariance testing is that "small" deviations from measurement invariance are not worth accounting for, and should be ignored.

Using fit measures instead of the null hypothesis chi-square test does recognize that some measurement differences are not worth accounting for, since they do not threaten substantive conclusions. Unfortunately, however, there is no mechanism in these omnibus tests that prevents large measurement differences from threatening the conclusions. It is possible for a misspecification to appear "small" relative to the worst case and degrees of freedom, while still being large enough to completely reverse substantive conclusions. Conversely, fit indices can also flag problems that are too small or unrelated to the substantive interpretations of interest. Practical examples of such "false negatives" and "false positives" from fit indices are given in Oberski (2014a).

An illustration

To try to resolve the above issues, this chapter discusses a complementary approach to measurement invariance evaluation: sensitivity analysis using the "Expected Parameter Change" in the parameter of interest, or EPC-interest (Oberski, 2014a). The EPC-interest in this context is a measure of the hypothetical change in a free parameter of substantive interest if a cross-group measurement invariance restriction were freed.

Examples of parameters of interest corresponding to the example analyses mentioned at the beginning of the chapter include the latent mean difference across countries in level of internal efficacy and the latent regression coefficient between efficacy in voting in different countries. Examples of cross-group measurement invariance restrictions include equality of slopes (e.g., parallel lines in Figure 1) and/or intercepts (e.g., equal line origins in Figure 1). For example, the EPC-interest could indicate how much change can be expected in the estimated cross-country difference in "internal efficacy" after allowing for a difference in intercepts for a certain item across countries. Thus, after fitting a measurement invariance model for a substantive purpose, the EPC-interest directly evaluates whether any potential violations of measurement invariance threaten that substantive purpose.

Data

To illustrate, I conducted a confirmatory factor analysis of three indicators of "internal political efficacy". These variables were measured in the European Social Survey (ESS), round seven (2014), and fifteen countries to be compared are included in this analysis. The items from the ESS questionnaire were:

- *actrolg*: How able do you think you are to take an active role in a group involved with political issues? Please use this card.
  (0 Not at all able – 10 completely able)
- *cptppol*: And using this card, how confident are you in your own ability to participate in politics? (0 Not at all confident – 10 extremely confident)
- *etapapl*: Using this card, how easy do you personally find it to take part in politics?
  (0 Not at all easy – 10 extremely easy)

These items were translated into the fifteen respective countries' official languages and administered to $n = 28{,}221$ respondents. The full dataset and questionnaire, as well as extensive details on the design, data collection, translation, and other processes that produced these data are all available publicly and freely at http://www.europeansocialsurvey.org/data/download.html?r=7. The total number of interviewed respondents, number of respondents used in the analysis, as well as means and standard deviations of the indicators per country used here are described in Table 1.

Table 1 shows some differences in the means of these indicators across countries. For example, Denmark (DK), Finland (FI), and Norway (NO) are in the top three highest scores for two of the three indicators (*actrolg* and *cptppol*) while Estonia (EE), Poland (PL), and Slovenia (SI) score lowest. These differences may be of interest to political scientists studying political participation. Here we will assume that what is of particular interest is not the country mean of each indicator, but rather an overall country mean of the latent variable "internal efficacy", which is assumed to underlie each of these indicators. The goal will then be to compare these latent means across the countries, and, ultimately, to interpret any differences found in substantive terms.

Table 1. Number of observations (N), means (M), and standard deviations (S) for all 15 countries in the European Social Survey, round 7 (2014). Countries are indicated with their ISO2 codes, see https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2.

	1. AT	2. BE	3. CH	4. CZ	5. DE	6. DK	7. EE	8. FI	9. FR	10. IE	11. NL	12. NO	13. PL	14. SE	15. SI
N	1795	1769	1532	2148	3045	1502	2051	2087	1917	2390	1919	1436	1615	1791	1224
actrolg															
M	2.33	2.48	2.57	2.27	2.62	3.14	1.86	3.27	2.48	2.50	2.35	3.14	2.01	2.96	1.61
S	(2.0)	(2.0)	(1.9)	(1.9)	(2.0)	(2.0)	(1.9)	(1.9)	(2.0)	(1.8)	(2.0)	(2.0)	(1.9)	(2.0)	(1.9)
cptppol															
M	2.86	2.69	3.38	2.07	3.37	3.00	1.92	2.81	2.95	2.65	2.51	3.27	1.97	3.09	1.74
S	(2.0)	(2.0)	(2.0)	(1.8)	(1.9)	(1.9)	(1.9)	(1.9)	(1.9)	(1.9)	(2.0)	(1.9)	(1.8)	(1.9)	(1.9)
etapapl															
M	2.47	2.73	3.19	2.72	2.82	3.09	1.93	3.17	2.76	2.38	2.77	2.93	1.77	3.00	1.95
S	(1.9)	(1.9)	(1.8)	(1.9)	(1.9)	(1.8)	(1.8)	(1.8)	(1.8)	(1.8)	(1.9)	(1.8)	(1.7)	(1.9)	(1.9)

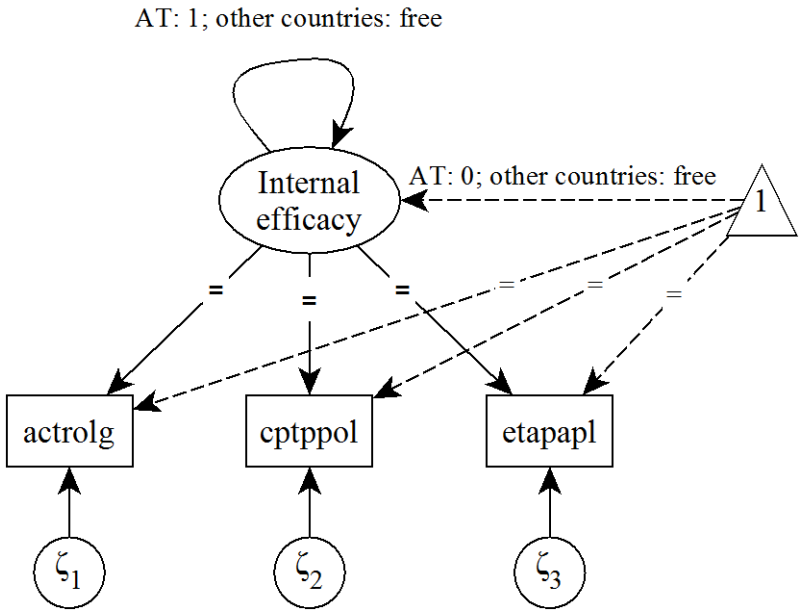

Figure 2. Scalar invariance model for internal efficacy in the European Social Survey, round 7 (2014). Dashed arrows from the triangle symbolize intercepts/means and the recurrent arrow on the latent variable indicates the variance. Equality over countries is indicated with the "=" symbol.

```
mod_effic_internal_scale <-

       "efin =~ actrolg + cptppol + etapapl

        efin ~~ c(var1, var2, var3, var4, var5, var6, var7, var8, var9, var10,

                  var11, var12, var13, var14, var15) * efin

        var1 == 1

        actrolg~~actrolg

        cptppol~~cptppol

        etapapl~~etapapl"


fit <- lavaan(mod_effic_internal_scale,

              data = ess7_sub, group = "cntry",

              auto.fix.first = FALSE, int.ov.free = TRUE, auto.var = FALSE,

              group.equal = c("intercepts", "loadings"))
```

Figure 3. R Lavaan code to fit the full scalar invariance model shown in Figure 2. The model is parameterized such that the first group (Austria, AT) is taken as a reference group, with latent mean zero and latent variance equal to unity.

Model

Figure 2 depicts the confirmatory factor analysis (CFA) model used to estimate the measurement relationships and latent means of substantive interest. The ellipses denote unobserved (latent) variables, while the rectangles denote the three observed indicators *actrolg*, *cptppol*, and *etapapl*. The dashed arrows from the triangle containing the number one are intercepts and latent means [1]. Finally, the arrow from the latent variable to itself denotes that variable's variance. The model in Figure 2 assumes zero residual correlations among the observed variables after accounting for the factor; in other words, the covariance among the three indicators is assumed to be due to a single underlying factor, named "internal political efficacy". We will assume that substantive interest focuses on the latent mean estimates for this factor, i.e. that the dashed path from the triangle to the latent variable is of primary interest.

The model shown in Figure 2 applies to all fifteen countries. Equality restrictions applied in the scalar invariance model, which identify the parameters of interest, are shown in Figure 2 as equality signs. The loadings and intercepts in Figure 2 are restricted to be equal across all fifteen countries. The latent variance is fixed to unity in the first (reference) group, Austria (AT), and left freely estimated in the other countries. Thus, the variance estimates in all countries are expressed as a ratio with the reference group. Similarly, the latent mean estimate of interest is fixed to zero in the reference group AT and freely estimated in all other countries, expressing the other latent mean estimates as differences with the reference group. While this parameterization is perhaps not the standard one familiar to most readers, it is mathematically equivalent to more traditional choices and holds the advantage of interpretability for our further analyses. Specifically, the latent mean estimates will have Austrian standard deviations as their unit of measurement. Thus, these estimates of interest are interpretable as effect sizes.

Three models with different levels of "measurement invariance" are commonly considered: configural, metric, and scalar invariance. These models form a hierarchy of consecutively more stringent equality restrictions. Configural invariance implies the same factor model holds across all groups, with identical patterns of zero cross-loadings and residual correlations. Metric invariance implies configural invariance, with loadings (slopes) that are equal across groups. Scalar invariance implies configural and metric invariance, and restricts measurement intercepts to be identical across groups. The model shown in Figure 2 is therefore a scalar invariance model.

I used the package `lavaan` 0.5-23.1019 (Rosseel, 2012) with `R` version 3.3.1 (R Core Team, 2016) to estimate the model in Figure 2 using the European Social Survey data. Figure 3 shows the `lavaan` model syntax and `R` code used to fit the Figure 2 model to these data. In the syntax, the line "`efin =~ actrolg + cptppol + etapapl`" indicates that the latent variable, now named `efin`, is measured by the three observed indicators. The three lines below that in the syntax ensure the first group's latent variance is fixed to unity, while the other groups' latent variances are freely estimated, as well as the residual variances of the observed variables. Below the `lavaan` syntax, the `R` code that fits the model is shown. Importantly, the argument `group = "cntry"` is used to indicate that a multiple group model is requested, and `group.equal = c("loadings", "intercepts")` leads to the scalar invariance model. For simplicity of exposition, we do not account for the ESS's complex sampling design here (see Oberski, 2014b for information on how the sampling design may be incorporated into the analysis), and use the default listwise deletion (complete case analysis) to deal with missing values [2]. For more information on `lavaan` and its syntax, the reader is referred to http://lavaan.ugent.be/.

---

[1] This follows the logic that a regression with the constant one as predictor will give as a regression coefficient "of" this constant the dependent variable's mean (when there are no other predictors, i.e., simple regression) or intercept (when there are, i.e. multiple regression).

[2] This could be remedied by specifying `missing = 'fiml'` as an argument to the `lavaan` function call.

Results

After fitting the model described in the previous section, we obtain (unstandardized) parameter estimates for the common loading and intercepts, shown in Table 2. These numbers determine the intercepts and slopes of measurement relations such as those shown for a different example in Figure 1. Model test statistics and fit indices are also produced for this model and described in Table 2. By most standards (e.g. Hu & Bentler, 1998), these indicate that the full scalar invariance model does not fit the data well.

To demonstrate the common practice of comparing different models in measurement invariance testing, I applied the `measurementInvariance` function from the `semTools` package (semTools Contributors, 2016), for which the results are shown in Table 3. The chi-square (χ^2) and improvement in chi-square between models ($\Delta\chi^2$) are shown in the last columns of Table 3. All improvements in chi-square are statistically significant. That is to say, the null hypothesis that the loadings and intercepts are exactly equal in the population is rejected. The CFI and RMSEA values in Table 3 can likewise be compared across models, as suggested by Chen (2007), who also provided recommended cutoff values for these differences (ΔCFI and ΔRMSEA). If these recommendations are followed, the researcher would conclude that neither scalar nor metric invariance holds (approximately) for these data. Finally, Akaike's Information Criterion (AIC) and Schwarz's (or "Bayesian") Information Criterion (BIC) are measures of the misfit of the model relative to its complexity (Raftery, 1995). These criteria select the model with the lowest value. AIC follows the fit measures and chi-square evaluation in selecting the configural invariance model, while BIC selects the metric invariance model.

In short, by all omnibus fit measures, the chi-square, and information criteria, the intercepts, and probably also the loadings, should not be constrained to be equal across countries. While this may indeed be realistic, it also means that the latent mean differences across countries cannot be estimated. After all, as the reader can easily verify using `lavaan`, the model that frees all intercepts as well as the latent means of interest is not identified. Similarly, if the loadings are freed across all countries, the variances of the latent variables are no longer identifiable. Note that restricting a loading to equal unity in each group would implicitly restrict that loading to be equal across groups and therefore obscure this fact (Hancock, Stapleton, & Arnold-Berkovits, 2009).

Table 2. Parameter estimates of the scalar invariance model. Chi-square (56 df): 1173.1 ($p < 10^{-6}$); CFI: 0.937; RMSEA: 0.132.

			95% C.I.	
	Est.	s.e.	Lower	Upper
Loadings				
actrolg	1.54	(0.04)	1.46	1.62
cptppol	1.71	(0.04)	1.62	1.80
etapapl	1.21	(0.03)	1.14	1.27
Intercepts				
actrolg	2.52	(0.05)	2.43	2.62
cptppol	2.69	(0.06)	2.58	2.80
etapapl	2.63	(0.04)	2.55	2.71

Table 3. Comparisons of models with different levels of measurement invariance.

	CFI	RMSEA	ΔCFI	ΔRMSEA		AIC	BIC	χ^2	$\Delta\chi^2$	df	$p(\chi^2)$
Configural	1	0			0	192,968	194,014	0			
Metric	0.993	0.063	0.007	0.063	28	193,068	193,896	155	155	28	$<10^{-6}$
Scalar	0.937	0.132	0.055	0.069	56	194,029	194,641	1,173	1,018	28	$<10^{-6}$

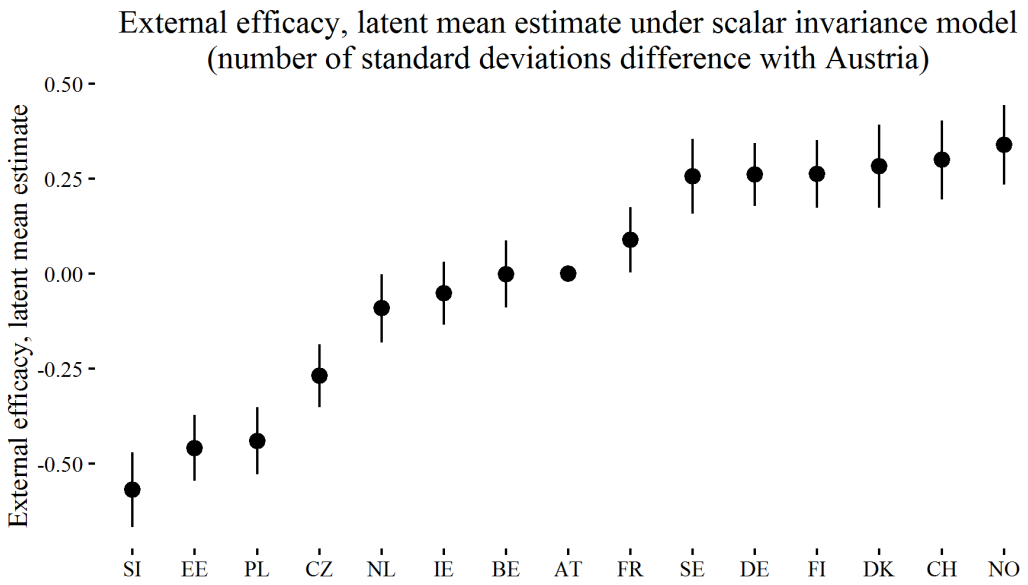

Figure 4. Parameters of interest: the latent country mean estimates with 95% C.I.'s, obtained under the (likely misspecified) full scalar invariance model. The reference country (Austria) is shown with its fixed estimate of zero and no C.I.

Remembering that the model was estimated with a substantive purpose in mind, the model fit evaluation above was done to protect the parameters of substantive interest from possible bias due to violations of the model assumptions. What would we conclude if we were to ignore these violations and plot the latent mean estimates under the (wrong) scalar invariance model? Figure 4 plots these parameters of interest, together with their 95% confidence intervals. As noted above, the points in Figure 4 can be interpreted as the number of (Austrian) standard deviations difference from Austria. The countries are ordered by the size of the estimate. For example, Slovenians (SI) are estimated to have a much lower average internal political efficacy score than the reference country, with an effect size difference of -0.5. Norwegians (NO) are on the top of the list, about a third of a standard deviation above the reference country but close to other, similar, countries.

Figure 4 demonstrates the type of result we might examine if the scalar invariance model were true. But the fit measures and chi-square have made us disbelieve that model. Thus, we are stuck between two difficult decisions: abandon the research of substantive interest to political science, or continue interpreting the estimates in Figure 4 at the risk that these may be biased, possibly affecting our conclusions. It is possible to search for a different, better-fitting model, but such a model will very likely exhibit the same problem: the assumption that even some measurement parameters are *exactly* equal across groups is usually false. Van de Schoot et al. (2013) described this dilemma as a choice between Scylla and Charybdis, two mythical monsters encountered by the Greek hero Odysseus. We now turn to sensitivity analysis as a possible method of navigation in the dangerous waters of cross-cultural analysis.

Sensitivity analysis

The traditional model fit evaluation approach to measurement invariance discussed above is focused on selecting a single "best" model, and basing all conclusions on this model. This procedure is unattractive when the "best" model may not be good enough, or when the best model does not allow for the conclusions of interest at all, as is the case when the parameters of interest are unidentifiable in it. An alternative approach to evaluating whether a model is adequate based on its fit to the data is to evaluate the extent to which that lack of fit may cause bias or, at least, different conclusions (Oberski, 2014a; Oberski & Vermunt, 2013). This idea was put forward in general terms by Leamer (1983) and is applied here to the case of measurement invariance. If relaxing some assumptions of measurement invariance can reverse the conclusions, then no definite conclusion is possible, one way or the other. On the other hand, if the conclusions are robust to violations of the measurement invariance restrictions then even a model that does not fit the data well can be usefully applied to study the parameters of substantive interest.

The most straightforward method of performing a sensitivity analysis in the context of measurement invariance testing is to fit all alternative possible models freeing one restriction at a time, including all combinations of such restrictions. Such processes could be automated using R scripts and the results summarized using plots and summary statistics. However, in spite of this possibility, it may be prohibitively complex for many applied researchers, especially since the combinatorial number of possible alternative models becomes exponentially larger with the number of items and groups.

An alternative to the manual sensitivity analysis method is to use an approximation to the change in substantive parameters of interest that would occur if an existing equality restriction were freed. This measure, the "EPC-interest", is similar to the well-known "expected parameter change" (here named "EPC-self" to avoid confusion), in that it quantifies a change due to freeing a certain restriction in the model. But whereas the EPC-self quantifies the changes in the restricted parameter *itself* when that parameter is freed, the EPC-interest quantifies the resulting change in the *parameter of interest*.

In our current example, the parameters of interest were chosen to be the latent mean estimates for the internal efficacy factor. The assumptions to be examined in a measurement invariance context are the cross-group equality restrictions on measurement parameters. The EPC-interest then asks questions such as, "how would the latent mean estimate in my model, such as those in Figure 4, change if the intercepts and slopes in Table 2 differed over countries?". In the current example, this question is asked once per intercept and slope restriction. That is, in the example analysis, which had three intercepts and three loadings restricted to be equal over the fifteen countries (groups), there are six sets of EPC-interest values indicating how the estimates in Figure 4 would change as a result.

Figure 5 shows R code that can be used in the current version of `lavaan` (0.5-23.1019 at the time of writing) to obtain the EPC-interest values for the cross-group equality restrictions on the intercept of the item *astrolg*. The resulting EPC-interest estimates are shown in Table 4. This table, therefore, gives an approximate indication of the amount of change in the latent mean estimates of substantive interest that can be expected after relaxing the assumption that the intercept `astrolg~1` is equal across groups. As can be seen in Table 4, the changes after relaxing this assumption are relatively small in absolute terms. If we follow Saris et al. (2009) in choosing 0.1 as an "important" substantive difference in effect size estimates, only the estimate for Finland is "importantly" affected. Because it can be difficult to see the effects of the measurement invariance assumptions quantified by EPC-interest in tables of numbers, Figure 6 plots these again for the results shown in Table 4, as well as for the five other sets of EPC-interest values.

```
# The code below works out which numbered restrictions correspond
#     to the first intercept (".p8."):
lav_test_all <- lavTestScore(fit, univariate = TRUE, epc = TRUE)
restricts <- which(lav_test_all$uni$lhs == ".p8.")


# Use the lavTestScore function to obtain the EPC-interest:
epc_p8 <- lavTestScore(fit, release = restricts, univariate = TRUE, epc = TRUE)


# Select only the changes in the latent mean estimates of substantive interest:
epc_interest <- subset(epc_p8$epc, lhs == 'efin' & op == "~1")
```

Figure 5. R lavaan code to obtain the EPC-interest values for freeing the cross-group equality restrictions on the first indicator of the latent efficacy factor.

Table 4. EPC-interest output from the code in Figure 5. Countries have been ordered by their estimate under the scalar invariance model (first row), also shown in Figure 4. The second row shows the country's EPC-interest when hypothetically freeing the cross-group equality restriction on the intercept of the item actrolg. The third row shows the corresponding hypothetical resulting new estimate that would be obtained if this restriction were freed.

	SI	PL	EE	CZ	IE	BE	AT	NL	FR	DE	FI	CH	SE	DK	NO
Estimate	-0.61	-0.47	-0.46	-0.31	-0.06	-0.01	0.00	0.01	0.06	0.30	0.36	0.41	0.42	0.43	0.51
EPC-interest	-0.04	-0.08	-0.06	-0.09	-0.06	-0.03	0.00	-0.03	0.00	0.03	-0.12	0.06	-0.05	-0.07	-0.05
Est. + EPC-interest	-0.65	-0.55	-0.52	-0.40	-0.11	-0.04	0.00	-0.02	0.07	0.33	0.24	0.46	0.37	0.36	0.46

Figure 6 contains six graphs. Each graph corresponds to the approximate change in the parameters of interest (the latent mean estimates) after freeing one set of equality restrictions. Since there are three intercepts restricted to equality across groups and three loadings restricted to equality across groups, this results in six graphs. Since a model with equal intercepts but differing loadings is highly implausible, in the lower row we consider the alternative estimates resulting from freeing the loading of each item *in addition to* its intercept. Each graph repeats the original estimates with their 95% C.I.'s from Figure 4 in gray, with the countries on the horizontal axes ordered by these current estimates ("Estimate" in Table 4). On top of these current estimates, the solid dots indicate what estimate might result if the corresponding set of equality restrictions were freed. Table 4 denoted these "Est. + EPC-interest", while `lavaan` lists them in its output as the "expected parameter values" (EPV). Black solid dots are EPV's that do not differ by more than 0.1 from their current estimate. Dark red dots differ by more than 0.1 and therefore signal sensitivity.

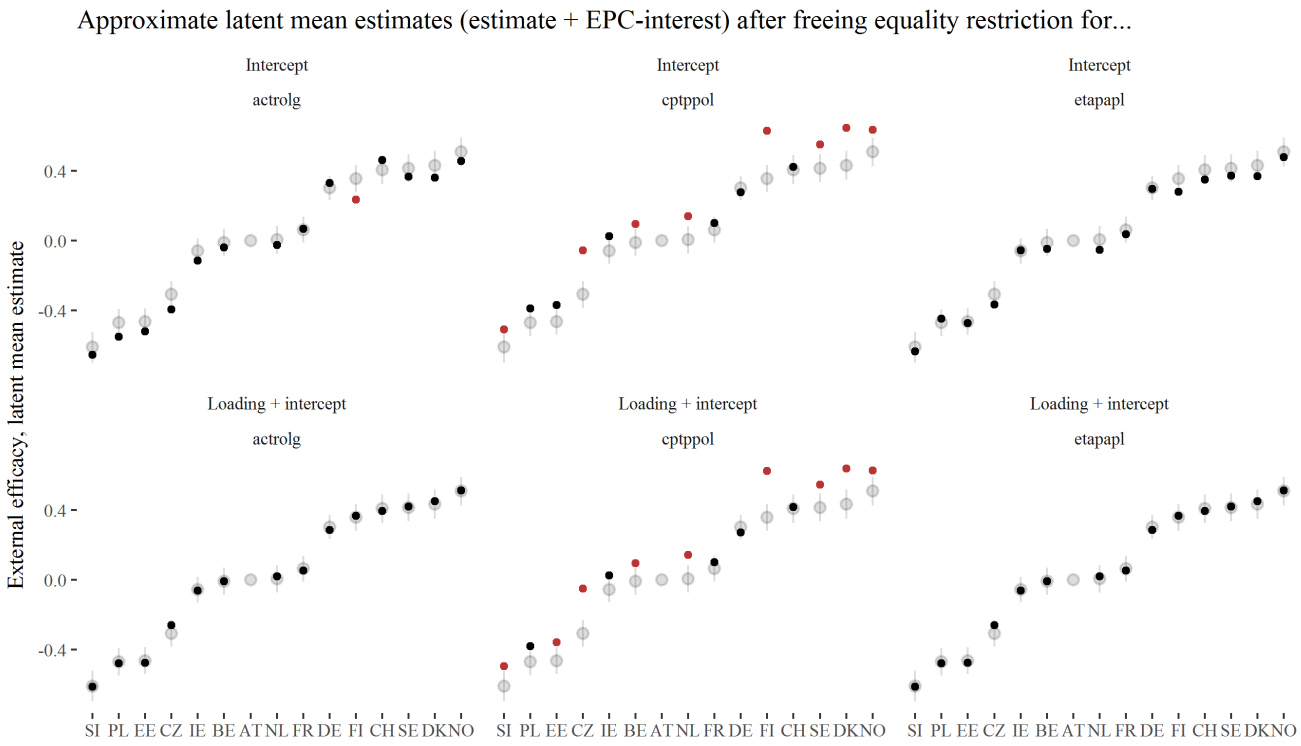

Figure 6. Sensitivity of the latent mean estimates to equality restrictions on intercepts and loadings. Gray intervals indicate the original estimates assuming full scalar invariance. Approximate latent mean estimates using EPC-interest are plotted in solid dots. Latent means that are sensitive to the equality restriction by more than 0.1 Austrian standard deviations are shown in red.

By the standards used here, it is clear that, even though the scalar invariance model fits the data very badly, the badness of fit that affects the conclusions of interest is almost entirely in the equality restrictions on the intercepts of the second indicator, *cptppol*. Therefore, the full scalar invariance model is clearly not robust to its assumptions. On the other hand, the researcher has been given the hope of finding an alternative model that frees some of the offending restrictions and is more robust. This alternative partial invariance model may also not fit the data, but its parameter estimates of substantive interest could be more robust to its misspecifications than those of the current model. Alternatively, it is possible that the partial invariance model freeing the equality restrictions on *cptppol* is also overly sensitive to violations of its remaining, identifying, cross-group equality restrictions. In that case freeing another indicator's intercept will lead to a model with untestable assumptions, whose sensitivity cannot be investigated; that case would therefore lead the researcher to conclude that the latent means are indeed not estimable in a way that is robust to model assumptions. The EPC-interest can be used to determine which of the two situations pertains in this case (code and data for the example analysis are available at http://).

Conclusion

This chapter has elaborated on sensitivity analysis as a method of dealing with the problem that measurement invariance is almost never true exactly. Fit measures, chi-square difference testing, and other procedures applied to measurement invariance testing do not account for the goal of the analysis and can therefore not guarantee that threats to the conclusions' validity have been ruled out, nor that rejected models were truly useless. Sensitivity analysis addresses these questions directly, by evaluating the (likely) impact of freeing cross-group equality restrictions on the parameters of substantive interest. The EPC-interest, an approximate measure of this impact, was introduced as a useful tool for sensitivity analyses, and the chapter showed how it can be calculated using standard open source software for structural equation modeling. The measure has also been implemented in the commercial latent variable model software Latent GOLD, which allows for categorical-data models such as IRT and latent class analysis (Vermunt & Magidson, 2013; Oberski & Vermunt, 2013). An example of such an analysis using a latent class multilevel model can be found in Oberski, Vermunt, & Moors (2015).

The usage of EPC-interest and sensitivity analysis also has disadvantages. It generally requires considerable effort on the part of the researcher, who is required to specify what in the analysis is of "substantive interest", and what would constitute a "serious change" in the conclusions. There are some obvious candidates for these choices, such as positive coefficients becoming negative or vice versa. However, there are also situations where the criteria to be applied may be less clear-cut. This means that the researcher will be forced to consider very carefully the meaning of measurement invariance in terms that are of substantive interest to him or her. Of course, that may be seen as an advantage by some, but certainly requires additional effort on the part of the researcher.

Another peculiar aspect of using sensitivity analysis in this manner is that the choice of outcome measure determines the results. It is no longer the case that measurement invariance is a property of the scale in itself. Instead, it becomes a joint property of the scale and the analysis of substantive interest. In the example analysis, for instance, the *rank order* of the countries was likely rather sensitive to small changes due to the fact that several countries' estimates were close together. This means that it will be difficult to find a misspecified model in which the misspecifications do not affect the rank order at all. In a sense, this observation merely reflects the law that, as more precise conclusions are drawn from the data, more stringent requirements on those data are needed. From that perspective, this aspect of sensitivity analysis simply reveals a truth about the limits to the conclusions that can be drawn from a given analysis.

Acknowledgements

Thanks are due to the anonymous reviewer for various suggestions that improved this chapter, especially the suggestion to consider the effect of loadings and intercepts jointly. Thanks are also due to Jeroen Vermunt and Yves Rosseel for their software implementations of the EPC-interest.

References

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

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

Coleman, K. M., & Davis, C. L. (1976). The Structural Context of Politics and Dimensions of Regime Performance:“ Their Importance for the Comparative Study of Political Efficacy.” *Comparative Political Studies*, *9*(2), 189.

Coromina, L., Saris, W. E., & Oberski, D. (2008). The Quality of the Measurement of Interest in the Political Issues presented in the Media in the ESS. *ASK. Research and Methods*, *17*, 7–38.

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In T. Teo & M. S. Khine (Eds.), *Structural equation modeling: Concepts and applications in educational research* (pp. 137–174). Rotterdam, The Netherlands: Sense Publishers.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.

Karp, J. A., & Banducci, S. A. (2008). Political efficacy and participation in twenty-seven democracies: How electoral systems shape political behaviour. *British Journal of Political Science*, *38*(2), 311–334.

Kim, U., Helgesen, G., & Ahn, B. M. (2002). Democracy, Trust, and Political Efficacy: Comparative Analysis of Danish and Korean Political Culture. *Applied Psychology*, *51*(2), 318–353.

Leamer, E. E. (1983). Let’s take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge.

Oberski, D. L. (2014a). Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models. *Political Analysis*, *22*(1), 45–60. https://doi.org/10.1093/pan/mpt014

Oberski, D. L. (2014b). lavaan.survey : An R Package for Complex Survey Analysis of Structural Equation Models. *Journal of Statistical Software*, *57*(1). https://doi.org/10.18637/jss.v057.i01

Oberski, D. L., & Vermunt, J. K. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 117–122.

Oberski, D. L., Vermunt, J. K., & Moors, G. B. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis*, *23*(4), 550–563.

R Core Team. (2016). R: A language and environment for statistical computing (Version 3.3.1, 2016-06-21, svn rev 70800). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111–163.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen & J. Scott Long (Eds.), *Testing structural equation models* (Vol. 154, p. 181). Sage.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

semTools Contributors. (2016). semTools: Useful tools for structural equation modeling (Version R package version 0.4-13). CRAN. Retrieved from http://cran.r-project.org/package=semTools

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company. Retrieved from http://psycnet.apa.org/psycinfo/2002-17373-000

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. *Belmont, MA: Statistical Innovations Inc*. Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf

Wright, J. D. (1981). Political disaffection. In *The handbook of political behavior* (pp. 1–79). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4684-3878-9_1