Published as:

Cieciuch, J., Davidov, E., Oberski, D.L., & Algesheimer, R. (2015). “Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples”. *European Political Science*, 14(4), pp 521–538. DOI: 10.1057/eps.2015.64.

Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples

Jan Cieciuch

jancieciuch@gmail.com

URPP Social Networks, University of Zurich, Switzerland

and Cardinal Stefan Wyszyński University in Warsaw, Poland

Eldad Davidov

Institute of Sociology, University of Zurich, Switzerland

Andreasstrasse 15; CH-8050 Zurich, Switzerland

Tel: +41 44 635 23 22; Fax: +41 44 635 23 99

Daniel L. Oberski

Department of Methodology and Statistics, Tilburg University, The Netherlands

René Algesheimer

Department of Business Administration, University of Zurich, Switzerland

Author Note

The work of the first, second, and fourth authors was supported by the University Research Priority Program (URPP) 'Social Networks', University of Zürich. The work of the third author was supported by the Netherlands Organization for Scientific Research (NWO) [Vici grant 453-10-002]. The second author would like to thank the EUROLAB, GESIS, Cologne, for their hospitality during work on this paper. The authors would also like to thank Lisa Trierweiler for the English proof of the manuscript.

Abstract

Political scientists often need to evaluate whether samples are comparable, for example, when analysing different countries or time points or when comparing data collected using different methods. A necessary condition for conducting such meaningful cross-group comparisons is the establishment of measurement invariance. One of the most frequently used procedures for establishing measurement invariance is the multigroup confirmatory factor analysis (MGCFA). This method was criticised in the literature because it may suggest that a model fits the data although it may contain serious misspecifications. We present an alternative method to test for measurement invariance using detection of local misspecifications and illustrate its use on two datasets assessing value priorities that are often analysed in political science and collected using paper-and-pencil and web modes of data collection.

Keywords: measurement invariance, detection for misspecification, multigroup confirmatory factor analysis (MGCFA), human values, statistical power, mode effects

Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples

Introduction

Political scientists frequently aim to evaluate whether the answers given by two or more groups of respondents are comparable. When they analyse international surveys such as the European Social Survey (ESS), the World Value Survey (WVS), the International Social Survey Program (ISSP), various election studies, or household panels, they must first establish that the measurement properties of survey questions are reliable, valid, and comparable across countries, time points and modes of data collection (Jowell et al., 2007; Oberski, 2012; Saris and Gallhofer, 2007).

Indeed, different modes of data collection may result in measurements that cannot be compared meaningfully (Gordoni et al., 2012; Révilla and Saris, 2012; Saris and Hagenaars 1997). The potential bias of the mode of data collection used is also acute in the analysis of international surveys because different modes of data collection are often used within and across countries to reduce costs (De Leeuw, 2005; Dillman et al., 2009). For example, the 2009 European Election Study used both telephone and face-to-face interviews[1]. Indeed, Podsakoff et al. (2012) showed that 18% to 32% of the total variance in measurement items is due to method bias and that mode of data collection is one of the most important predictors of item validities and reliabilities. Thus, using differing modes of data collection will impact items' reliability and validity and may constitute a major source of measurement bias (for further discussion, see Podsakoff et al., 2012).

For these reasons, a practice of evaluating measurement comparability across samples known as "measurement invariance testing" has emerged in the literature (for overviews, see Millsap and Everson, 1993; Schmitt and Kuljanin, 2008; Vandenberg and Lance, 2000).

Various methods have been proposed in the literature to test for measurement invariance. One of the most frequently used procedures for establishing measurement invariance is multigroup confirmatory factor analysis (MGCFA). However, this method has been increasingly criticised because it may suggest that a model fits the data although it may contain serious misspecifications. In other words, such a model may suggest that scores are comparable across samples whereas in reality scores may still be biased and noncomparable (Saris et al., 2009). We present an alternative method to test for measurement invariance using detection of local misspecifications based on a study by Saris et al, (2009), and we illustrate its use in two datasets on value priorities collected using paper-and-pencil and web modes of data collection.

Our contributions are twofold: First, we extend the method proposed by Saris et al. (2009) of model evaluation to the evaluation of models testing for measurement invariance; second, we present the method and show how to apply it to testing for measurement invariance across different samples; third, we demonstrate its use on real data and test for invariance across two samples collected using different methods of data collection. For the illustration we use two samples using different modes of data collection to measure human values. Human values have been increasingly used in political science to explain political attitudes, voting behaviour or political orientation (see, e.g., Caprara et al., 2006; Piurko et al., 2011; Schwartz et al., 2014; Vecchione et al., 2015). In the concluding section we explain how the method may be applied also to assess the comparability of scores across different cultures, language groups, countries, or other groups. We begin by first briefly explaining the importance of measurement invariance and how it can be tested.

Measurement Invariance

The Importance of Measurement Invariance

Measurement invariance is a psychometric measurement property. Measurement invariance affirms that a questionnaire does indeed measure the same construct in the same way in various groups or at various time points or under different conditions (Chen, 2008; Davidov, Meuleman, Cieciuch, Schmidt, Billiet, 2014; Marsh et al., 2010; Meredith, 1993; Millsap, 2011; Steenkamp and Baumgartner, 1998; Van de Vijver and Poortinga, 1997; Vandenberg, 2002; Vandenberg & Lance, 2000).

The assumption that a questionnaire is measurement invariant is a precondition for 1) a meaningful comparison of data across groups, time points, or conditions and 2) for pooling data collected in different groups for further analysis. Measurement invariance does not imply that there are no differences between the groups with regard to the measured construct. Instead, measurement invariance simply implies that a measured construct is comparable across various groups or modes of data collection and that its score is not biased by different parameters of the items measuring it. Therefore, when measurement invariance is established, the scores of the measurement can be meaningfully compared across groups and interpreted as being similar or different.

How Can Measurement Invariance Be Established?

There are many procedures for measurement invariance testing across groups (Chen, 2008; Davidov et al., 2014; Vandenberg and Lance, 2000). The most widely used method is multigroup confirmatory factor analysis (MGCFA; Jöreskog, 1971). This method involves setting cross-group constraints on parameters and comparing more restricted models with less restricted models (Byrne, 2004; Byrne et al., 1989; Byrne and Stewart, 2006; Davidov et al., 2014; Meredith, 1993; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000).

Let us assume that there is a questionnaire *Q*, consisting of six items, X_1 through X_6. The items are observed variables that serve as indicators measuring two latent variables, η_1

and η_2. Items X_1, X_2, and X_3 are indicators measuring latent variable η_1, and items X_4, X_5, and X_6 are indicators measuring latent variable η_2. The factor loadings of the indicators are denoted as λ (lambda), the indicator intercepts as τ (tau), and the variances of the measurement errors as ϑ (theta). The two latent variables (η_1 and η_2) are correlated and the variances of the latent variables are constrained to 1 for identification. It is also possible to identify the model by alternatively constraining one of the factor loadings for each latent variable to 1 (Bollen, 1989) with the so-called marker-variable method (Little et al., 2006). Data are collected using questionnaire *Q* for two groups or conditions, *A* and *B*. Figure 1 presents the measurement model for the two groups.

Figure 1 about here

Levels of Measurement Invariance: Configural, Metric, and Scalar

One can differentiate among several levels of measurement invariance. Each level is defined by the parameters constrained to be equal across groups. The first and lowest level of measurement invariance is called configural invariance (Horn and McArdle, 1992; Meredith, 1993; Vandenberg and Lance, 2000). Configural invariance requires that the same latent variables are measured by the same items for all groups. The model parameters are estimated for all groups simultaneously without any equality constraints. The fit of the model being tested provides the baseline against which the models testing for higher levels of measurement invariance are analysed.

The second level is called metric invariance (Horn and McArdle, 1992; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000) or weak measurement invariance (Marsh et al., 2010; Meredith, 1993). Metric invariance is tested by constraining the factor loadings between the observed items and the corresponding latent variable to be equal across the compared groups (Vandenberg and Lance, 2000). If metric measurement invariance is established one may assume that the meaning of the latent variable in both groups is the

same, although there is still a lack of certainty as to whether the construct is being measured in the same way in both groups.

A third and higher level of measurement invariance is called scalar invariance (Vandenberg and Lance, 2000) or strong measurement invariance (Marsh et al., 2010; Meredith, 1993). Scalar measurement invariance is tested by constraining not only the factor loadings but also the indicator intercepts to be equal across groups (Vandenberg and Lance, 2000). Intercepts and means are added into the model testing for scalar invariance. This type of analysis is referred to as the means and covariance structure analysis (MACS). If scalar invariance is established, one may assume that respondents use the scale in the same way in each group; thus, it implies that the same construct (metric invariance) is measured in the same way (scalar invariance). Steenkamp and Baumgartner (1998) called scalar invariance "score equivalence", since any difference between groups in observed scores corresponds to an equal difference in the latent variable scores. Table 1 summarises the different levels of measurement invariance and the comparisons that they allow if established.

Table 1 about here

The decision regarding which level of measurement invariance should be tested and established depends on the research goals. Configural invariance enables only an overall test of the similarity of the measurement models of the groups under study. In contrast, metric measurement invariance also allows for the comparison of unstandardised relationships between the constructs across groups. For example, this level of invariance allows a comparison of the covariance between $\eta 1$ and $\eta 2$ across groups A and B and the comparison of the strength of association of an external variable *V* with $\eta 1$ (or $\eta 2$) across the groups. Scalar invariance is needed to conduct meaningful comparisons of latent means across groups. Therefore, if a researcher is interested in whether the latent mean of variable $\eta 1$ (or $\eta 2$) is larger in group A or group B, scalar invariance should be established beforehand.[2] The

assessment whether measurement invariance at a certain level is given can be performed based on an evaluation of the fit of a model with the corresponding constraints (Chen, 2007; Davidov et al., 2014). This model evaluation has typically relied on global fit measures. In the current study we present another approach to evaluate the model based on the presence or absence of local misspecifications (Saris et al., 2009).

The Global Fit Measures Approach to Assessing Measurement Invariance

Historically, the first global fit measure in structural equation modelling (SEM) was the χ^2 (chi-square). According to Jöreskog (1969), the χ^2 value aided in freeing confirmatory factor analysis (CFA) of many subjective decisions that had to be made in exploratory factor analysis. As stated by Hu and Bentler (1995), the subjective judgement was replaced by an objective test of differences between two matrices: the observed and the hypothesised one in the model. Unfortunately, χ^2 is not free of problems, which have been widely recognised and discussed in the literature (e.g., Bentler and Bonett, 1980; Hu and Bentler, 1998; Hu et al., 1992; Kaplan, 1990). Two main problems were identified. The first is that the models tested in CFA are always only approximations of reality. Therefore, using χ^2 to test the hypothesis that the observed covariance matrix equals the hypothesised matrix is an unnecessary strong assumption (Jöreskog, 1978). The second problem is that χ^2 is sensitive to various characteristics of the tested model that are irrelevant to the possible misspecification. The most known case is sample size. As Bentler and Bonett (1980) pointed out more than 30 years ago, "in large samples, virtually any model tends to be rejected as inadequate" (p. 588). High power leads to the undesirable situation that substantively unimportant misspecifications can lead to rejection of an otherwise "closely fitting" model.

To resolve this problem, many fit indexes have been developed (Hu and Bentler, 1995; Marsh et al., 2005), and various recommendations for cut-off criteria for these model fit coefficients have been proposed. Three popular indexes, the root mean square error of

approximation (RMSEA), the comparative fit index (CFI), and the standardised root mean square residual (SRMR), are reported quite frequently in the literature. The RMSEA reflects the degree to which a researcher's model reasonably fits the population covariance matrix while considering the degrees of freedom and the sample size (Brown, 2006). This index is a parsimony-adjusted index that favours simpler models. The probability of close fit (Pclose) value indicates the probability that RMSEA is below 0.05. When the RMSEA value is lower than 0.05, the model can be assumed to perform very well (Brown and Cudeck, 1993). When the RMSEA value is 0.08 or lower, it can be assumed that the model performs reasonably well (Hu and Bentler, 1999; Marsh et al., 2004). The CFI compares the fit of a researcher's model to a more restricted baseline model. CFI values between 0.90 and 0.95 or larger indicate an acceptable model fit (Hu and Bentler, 1999). The SRMR compares the sample variances and covariances with the estimated ones. When the SRMR value is lower than 0.05, the model can be assumed to perform very well, and when it is lower than 0.08, the model can be assumed to perform reasonably well (Hu and Bentler, 1999; Marsh et al., 2004).

To assess whether a given level of measurement invariance is established, global fit measures are *compared* between the more and less constrained models. If the *change* in model fit is smaller than the criteria proposed in the literature, measurement invariance for that level is established. Chen (2007) proposed the use of cut-off criteria in deciding whether the fit of a more restrictive model has deteriorated significantly. According to Chen (2007), if the sample size is larger than 300, metric noninvariance is indicated by a change in CFI larger than .01 when supplemented by a change in RMSEA larger than .015 or a change in SRMR larger than .03 compared with the configural invariance model. Regarding scalar invariance, noninvariance is evidenced by a change in CFI larger than .01 when supplemented by a change in RMSEA larger than .015 or a change in SRMR larger than .01 compared with the metric invariance model.[3]

The Local Misspecification Approach

Saris et al. (2009) argued that because the power of both χ^2 and other fit indices to detect misspecifications are affected not only by the misspecification size but also by other characteristics of the models (e.g., sample size, model size, number of indicators, reliability of the indicators, size of parameters; see Cohen, 1988), relying on the global fit of the entire model (or on global fit differences) may lead to a wrong decision. The alternative approach to making decisions regarding model correctness is proposed by Saris et al. (2009) and involves investigation into whether specific misspecifications are present in the model. The correct model should not contain any relevant misspecifications. Thus, it is possible that, according to the common global fit criteria, one could accept a model that in reality contains serious misspecifications and should be rejected. It is also possible that the global fit measures will recommend the rejection of a model that does not contain any relevant misspecification and may actually be accepted (Saris et al., 2009).

Models are however simplifications of reality; therefore, they always contain misspecifications (Brown and Cudeck, 1993; MacCallum et al., 1996). Thus, the important question deals with how large a misspecification the researcher is willing to accept in a given model (Saris et al., 2009). Saris et al.'s (2009) approach enables detection of the misspecifications according to the misspecification size defined by the researcher. According to Saris et al. (1987), estimates of the misspecifications can be obtained using a combination of the expected parameter change (EPC) and the modification index (MI), which are usually provided by SEM software packages. The MI (Sörbom, 1989) provides information on the minimal decrease in the χ^2 of a model when a given constraint is released. Decreasing χ^2 leads to an improvement of the model. EPC provides a prediction of the change of the given parameter when it is released (Saris et al., 1987). Thus, EPC provides a direct estimate of the

size of the misspecification, whereas the MI provides a significance test for the estimated misspecification (Saris et al., 1987).

Neither the EPC nor the MI is error-free. The EPC estimation is problematic because of sampling fluctuations that may influence it. In addition, the value of the EPC depends also on other misspecifications in the model. The MI, similar to χ^2, depends on various characteristics of the model and the sample (Saris et al., 2009). To resolve this problem, Saris et al. (2009) introduced the standard error of the EPC and the power of the MI test. According to Saris et al. (1987), both the standard error of the EPC and the power can be estimated based on the MI and EPC.

Decision Rules in the Local Misspecification Approach

Before making the decision about whether the EPC is a significant indicator of relevant misspecification, the power of the MI test (which is not provided by the SEM program) should first be taken into consideration. The power can be calculated, however, using the EPC and the MI with the Jrule program developed by Oberski (2009) and van der Veld and colleagues (van der Veld et al., 2008; Saris et al., 2009).[4]

Before looking for local misspecifications, the researcher should make two decisions. The first requires specifying what size of misspecification should be tolerated. Keeping in mind that misspecifications are fixed to zero although they are not zero in reality, the question is how large is the difference between the 'zero' specified in the model and the real parameter (or parameter difference) that will be treated as a misspecification. Saris et al. (2009) formulated suggestions that should be used with caution: Deviations larger than .4 for cross-loadings and deviations larger than .1 for other parameters (differences in factor loadings or intercepts across groups) may be treated as serious misspecifications, and smaller ones may be ignored. The logic behind the first suggestion is that one usually considers factor loadings of .4 and higher as substantial (Brown, 2006). The logic behind the second

suggestion depends on the length of the scale used. Obviously, both suggestions are somewhat arbitrary, and a researcher may analyse the model using higher or lower misspecifications.[5] The second question that should be addressed by a researcher deals with the size of power that is considered high enough to detect the defined size of misspecification. Saris et al. (2009) suggest the value of .75 as a threshold of high power[6] . Both of these points (misspecification size and the size of power to detect potential misspecifications) must be indicated by the researcher in advance in the Jrule program (Oberski, 2009; Saris et al., 2009).

The Jrule program performs two operations: The first is to calculate the power of the test based on the MI and EPC (these may be provided by the SEM software, e.g., Mplus: Muthèn and Muthèn, 1998-2012, or Lisrel: Jöreskog and Sörbom, 2001). The second is to suggest, for each parameter, whether it is misspecified according to the decision criteria outlined in Table 2, taking into consideration the size of misspecification and the required power as defined by the researcher. Jrule calculates the actual power of the test of each parameter.

Table 2 about here

When the MI is significant and the power of the MI test is low, the researcher can conclude that there is a misspecification at the predefined level. When the MI is not significant and the power is high, the decision is also simple – there is no misspecification. Two other cases are more complicated. When the MI is significant and the power of the MI test is high, one does not know whether the significance is a result of a high or serious misspecification or whether it is significant just because the power of the test is high. In this case, Saris et al. (2009) and van der Veld and Saris (2011) suggest examining the substantive relevance of the EPC. The last case where the power is low and the MI is not significant is inconclusive, because although the MI is not significant, there is not enough information to

conclude whether this is a result of the low power or simply because there is no misspecification.

When measurement invariance is tested, several misspecifications are possible. First, the program may suggest that there are *cross-loadings* that should be accounted for. If the program proposes to consider cross-loadings (listed in the MI output of Mplus as 'by' statements) only for some but not all of the groups, then even configural measurement invariance is threatened.[7] Second, the program may suggest in the MI output (listed in the MI output of Mplus also as 'by' statements) that certain equality constraints on the *factor loadings* of the indicators measuring the latent variables are misspecified (i.e., not equal). Such a statement should be interpreted as an indication that the difference between the factor loadings across groups is larger than zero and that metric invariance is not supported by the data. Finally, the program may suggest that the *intercept parameters* equality constraints between groups are misspecified (listed in the MI output of Mplus as item names). Such a statement should be interpreted as an indication that the difference between the intercepts across groups is larger than zero and that scalar invariance is not supported by the data. Below we illustrate how the method may be applied.

Illustration

The current illustration tests the measurement invariance properties of human values (Schwartz, 1992) data collected using the PVQ-40 questionnaire (Schwartz et al., 2001). The Schwartz concept and measurement of values was quite often used in the political science literature (Caprara et al., 2006; Piurko et al., 2011; Schwartz et al., 2014; Vecchione et al., 2015). A previous study (Davidov and Depner, 2009) that tested for the measurement invariance properties of human values using the PVQ-21 questionnaire in the European Social Survey, showed that the values measured by the PVQ-21 are configural, metric, and

scalar invariant across online and paper-and-pencil conditions. Although the PVQ-21 was developed to measure ten values, Davidov and Depner (2009) tested the invariance properties for only seven values, because some values have to by unified in the CFA model. In the current study we use the full version of the questionnaire (40 items) that enables us to test for the invariance properties of more narrowly defined values. Based on (1) the reinterpretation of the questionnaire proposed by Cieciuch and colleagues (Cieciuch and Schwartz, 2012; Cieciuch et al., 2013), which was built on the refined version of the value theory (Schwartz et al., 2012), and (2) on the magnifying glass strategy proposed by Cieciuch and Davidov (2012), we test for the measurement invariance of sixteen more narrowly defined values measured by PVQ-40. Thus, we conduct the test for each higher-order value separately. The definitions of the ten values and the sixteen more narrowly defined values are presented in Table 3. For the illustration we used the Mplus 7 program (Muthén and Muthén, 1998-2012).

Method

Sample. The study was conducted on a group of N = 1,256 Polish adults. The online group consisted of n = 627 participants aged 18-69 years (58.7% female). The paper-and-pencil sample in the analysis contained n = 629 respondents aged 18-67 years (57.1% female). The data were collected by research assistants at the Cardinal Stefan Wyszyński University in Warsaw, Poland. For the data collection they asked their peers and acquaintances to complete the questionnaire. The online and paper-and-pencil samples were recruited in a similar way, and therefore there were no systematic differences across the samples in this regard. Participation was voluntary, and anonymity was guaranteed.[8]

Questionnaire. We used the Polish version of the PVQ-40 (Cieciuch and Schwartz, 2012). This questionnaire includes 40 short verbal portraits of different people who are gender matched to the respondent. Each portrait describes a person's goals, aspirations, or desires, which point implicitly to the importance of one of the ten basic values in the original

theory. For example, "Thinking up new ideas and being creative is important to her. She likes to do things in her own original way" describes a person to whom self-direction values are important. For each portrait, respondents answer "How much like you is this person?" on a scale of 1 (*not like me at all*) to 6 (*very much like me*).

Table 3 about here

Results

First, we ran a CFA analysis (Bollen, 1989; Brown, 2006) for each group and higher-order value separately. Thus, we created a separate CFA model for each higher-order value (four in total) where we allowed the values to correlate with each other. The items and the corresponding values are listed in the fourth column of Table 3. We were forced to exclude one facet of conformity (conformity-rules) and one facet of tradition (humility) from further analysis because they produced overly high correlations with other latent variables that made it impossible to differentiate between them. We tested four models in each sample: self-transcendence (with universalism-concern, universalism-nature, and benevolence correlating with each other), self-enhancement (with achievement-ambition, achievement-success, hedonism, power), openness to change (with self-direction-thought, self-direction-action, and stimulation), and conservation (with security societal, security personal, tradition [without humility], and conformity-interpersonal) with items as listed on Table 3.[9] Each of these four models resulted in satisfactory global model fit coefficients in both samples.[10]

Next, we performed MGCFA across samples for each of the four higher-order values beginning with configural invariance, then turning to metric invariance, and finally to scalar invariance sequentially. Global fit measures for the configural, metric and scalar levels of measurement invariance are presented in Table 4.

Table 4 about here

As Table 4 demonstrates, according to the cut-off criteria that are described above and customarily used, the scalar invariance models for each higher-order value fit the data well (CFI values above .90 and RMSEA values below .08; change in CFI and RMSEA below the cut-off criteria).

Next, we turned to the detection of local misspecification using the Jrule program. We looked for misspecifications larger than .4 for the "by" statements (latent variable by item, i.e., detection of the presence of cross-loadings) of parameters not present in the model, which correspond to misspecifications at the configural level, because these misspecifications imply the erroneous omission of cross-loadings. In the next step we looked for two types of misspecifications simultaneously: First, we looked for misspecifications larger than .1 for the "by" statements (latent variable by item) of parameters already estimated in the model, which correspond to misspecifications at the metric level of invariance, because the misspecifications concern incorrect equality constraints of *factor loadings* across groups. Second, we looked for misspecifications larger than .1 for item intercepts (the names of these items are indicated in the Mplus output) that correspond to misspecifications at the scalar level, because the misspecifications concern incorrect equality constraints of *intercepts* across groups.

Table 5 presents the results. The table contains the information provided by Mplus (MI and EPC) and by the Jrule program (power and decision). The power level required in Jrule was predefined to be 0.75.

Table 5 about here

As Table 5 demonstrates, there were no serious misspecifications at the configural level because the power was high and all misspecifications were below .4. There were also no serious misspecifications at the metric level because Mplus did not identify any misspecification for the "by" statements (latent variable by item) of parameters already

estimated in the model. In other words, all the cross-sample equality constraints on factor loadings between the values and their corresponding items were supported by the data and metric invariance could be established.

At the scalar level, the equality constraints of intercepts for both items se14 and se35 measuring societal security were misspecified. In other words, the intercepts of the two items considerably differed across the samples. This implied that the scores of the latent variable societal security could not be meaningfully compared across the samples. Different intercepts do not allow to compare the value's scores across samples.[11] Thus, although the global model fit presented in Table 4 for the corresponding scalar invariance model was acceptable, the precise detection of misspecification using Jrule led to the conclusion that one of the values of conservation, societal security, is not scalar measurement invariant while all other values are.

Conclusion

Measurement invariance is necessary for conducting meaningful cross-group comparisons. If measurement invariance is not established, interpretations of comparisons between groups are problematic (Chen, 2008; Davidov et al., 2014). Two main dangers arise from the use of noninvariant measurement instruments. The first danger is that different constructs may be measured across the various groups with the same measurement instrument. Horn and McArdle (1992) metaphorically compare such a case to a comparison between apples and oranges, and Chen (2008) describes it as a comparison between chopsticks and forks. The second danger is that despite measuring the same construct, it might have a different meaning across groups or respondents might respond to the questions measuring it in a different way, which also disrupts the ability to make meaningful comparisons. If a measurement is noninvariant, it is possible that differences that are found between groups do not correspond to real differences, or conversely, that the real differences are obscured by the noninvariant

measurements. Indeed, several scholars (see, e.g., King et al., 2004) have reminded us that measurement invariance cannot be taken for granted and has to be empirically tested.

There are many procedures for measurement invariance testing across groups. The most widely used method is multigroup confirmatory factor analysis. In this method, specific measurement parameters are constrained to be equal across groups, and researchers may rely on global fit measures to assess if the constraints are tenable. However, the literature raises criticisms about the use of global fit measures for assessing model fit (Saris et al., 2009). Instead, it suggests relying on MI, EPC, and the power of the test to ascertain if a model fits the data. In this study, we present this alternative method to assess whether measurement invariance is supported by the data. Furthermore, we illustrate its use for testing for measurement invariance across two samples.

In our illustration we demonstrate how to apply this procedure for establishing measurement invariance across human values measurements collected using two modes of data collection: online and paper-and-pencil. It turns out that our value measurements are highly invariant across online and offline modes of data collection. Whereas the global fit measures suggest that all values are invariant, the detection of misspecification procedure identifies one value, societal security, which is metric but not scalar measurement invariant. In other words, whereas associations between societal security and other theoretical constructs of interest may be compared across samples (as metric invariance was supported by the data), its means may not be comparable (as scalar invariance was not supported by the data). Thus, the method of misspecification detection is more sensible for identifying violations of invariance also when the global fit measures are satisfactory.

It should be noted, however, that the presented misspecification detection method does not provide guidance to determine which misspecifications may be tolerated and which may not. Such guidelines are neither provided for tests of metric invariance nor for tests of

scalar invariance. Thus, researchers need to decide by themselves which misspecifications they wish to tolerate and which they do not. Recently, Oberski (2014) and Meuleman (2012) provided additional guidelines for making this decision. Further studies should provide more precise guidelines as to which misspecifications may endanger the meaningfulness of cross-group comparisons.

Indeed, the method to detect for misspecifications illustrated in the current study on online and paper-and-pencil data may also be applied by political scientists when testing for measurement invariance across various samples such as cultural groups, time points, language groups, or any other groups that they wish to compare. The decision which groups to compare depends on the substantive research questions. Testing for measurement invariance may provide scholars with more precise indications about the extent to which their data are comparable after all. Findings of measurement invariance will guarantee that substantive comparisons across groups may be conducted with confidence.

References

Asparouhov, T. and Muthén, B.O. (2013) *Multiple group factor analysis alignment*. Mplus Web Note No. 18, Version 3, available at http://www.statmodel.com/examples/webnote.shtml, accessed 23 August 2013.

Bentler, P.M. and Bonett, D.G. (1980) 'Significance tests and goodness of fit in the analysis of covariance structures', *Psychological Bulletin* 88(3): 588-606. doi:10.1037//0033-2909.88.3.588

Bollen, K.A. (1989) *Structural Equation Modeling with Latent Variables*, New York: Wiley.

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

Brown, T.A. and Cudeck, R. (1993) Alternative Ways of Assessing Model Fit, in K.A. Bollen and J.S. Long (eds.) *Testing Structural Equation Models*, Newbury Park, CA: Sage, pp. 136-162.

Byrne, B.M. (2004) 'Testing for multigroup invariance using AMOS graphics: a road less traveled', *Structural Equation Modeling* 11(2): 272-300. doi:10.1207/s15328007sem1102_8

Byrne, B.M., Shavelson, R.J. and Muthén, B.O. (1989) 'Testing for the equivalence of factor covariance and mean structures - the issue of partial measurement invariance', *Psychological Bulletin* 105(3): 456-466. doi:10.1037/0033-2909.105.3.456

Byrne, B. M. and Stewart, S.M. (2006) 'The MACS approach to testing for multigroup invariance of a second-order structure: a walk through the process', *Structural Equation Modeling* 13(2): 287-321. doi:10.1207/s15328007sem1302_7

Caprara, G. V., Schwartz, S. H., Capanna, C., Vecchione, M. and Barbaranelli, C. (2006) 'Personality and politics: Values, traits, and political choice', *Political Psychology* 27: 1-28. doi: 10.1111/j.1467-9221.2006.00447.x

Chen, F.F. (2007) ‘Sensitivity of goodness of fit indexes to lack of measurement invariance’, *Structural Equation Modeling* 14(3): 464-504. doi:10.1177/0734282911406661

Chen, F.F. (2008) 'What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research', *Journal of Personality and Social Psychology* 95(5): 1005-1018. doi:10.1037/a0013193

Cieciuch, J. and Davidov, E. (2012) ‘A comparison of the invariance properties of the PVQ-40 and the PVQ-21 to measure human values across German and Polish samples’, *Survey Research Methods* 6(1): 37-48.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R. and Schwartz, S.H. (2014) ‘Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values’, *Frontiers in Psychology* 5:982. doi:10.3389/fpsyg.2014.00982

Cieciuch, J. and Schwartz, S.H. (2012) ‘The number of distinct basic values and their structure assessed by PVQ-40’, *Journal of Personality Assessment* 94(3): 321-328. doi:10.1080/00223891.2012.655817

Cieciuch, J., Schwartz, S.H. and Vecchione, M. (2013) ‘Applying the refined values theory to past data: what can researchers gain?’, *Journal of Cross-Cultural Psychology* 44(8): 1215-1234. doi:10.1177/0022022113487076

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

Cohen, J. (1992) ‘A power primer’, *Psychological Bulletin* 112(1): 155-159. doi:10.1037/0033-2909.112.1.155

Davidov, E. and Depner, F. (2009) ‘Testing for measurement equivalence of human values across online and paper-and-pencil surveys’, *Quality & Quantity* 45(2): 375-390. doi:10.1007/s11135-009-9297-9

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. and Billiet, J. (2014) ‘Measurement equivalence in cross-national research.’, *Annual Review of Sociology* 40: 55-75. doi: 10.1146/annurev-soc-071913-043137

de Beuckelaer, A. and Swinnen, G. (2011) Biased Latent Variable Mean Comparisons Due to Measurement Noninvariance: A Simulation Study, in E. Davidov, P. Schmidt and J. Billiet (eds.) *Cross-Cultural Research: Methods and Applications*, New York: Routledge, pp. 117-147.

De Leeuw, E.D. (2005) 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics* 21(5): 233-255.

Dillman, D.A., Smyth, J.D. and Christian, L.M. (2009) *Internet, Mail and Mixed-Mode Surveys. The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

Gordoni, G., Schmidt, P. and Gordoni, Y. (2012) 'Measurement invariance across face-to-face and telephone modes: the case of minority-status collectivistic-oriented groups', *International Journal of Public Opinion Research* 24*(*2): 185-207. doi:10.1093/ijpor/edq054

Horn, J.L. and McArdle, J.J. (1992) 'A practical and theoretical guide to measurement invariance in aging research', *Experimental Aging Research* 18(3-4): 117-144. doi:10.1080/03610739208253916

Hu, L.T. and Bentler, P.M. (1995) Evaluating Model Fit, in R. Hoyle (ed.) *Structural Equation Modeling: Issues, Concepts, and Applications*, Newbury Park, CA: Sage, pp. 76-99.

Hu, L.T. and Bentler, P.M. (1998) 'Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification', *Psychological Methods* 3(4): 424-453. doi:10.1037/1082-989x.3.4.424

Hu, L.T. and Bentler, P.M. (1999) 'Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives', *Structural Equation Modeling* 66(1): 1-55. doi:10.1080/10705519909540118

Hu, L.T., Bentler, P.M. and Kano, Y. (1992) 'Can test statistics in covariance structure-analysis be trusted?', *Psychological Bulletin* 112(2): 351-362. doi:10.1037/0033-2909.112.2.351

Jowell, R., Roberts, C., Fitzgerald, R. and Eva, G. (2007) *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey,* London: Sage

Jöreskog, K.G. (1969) 'A general approach to confirmatory maximum likelihood factor analysis', *Psychometrika* 34(2): 183-202. doi:10.1007/bf02289343

Jöreskog, K.G. (1971) 'Simultaneous factor analysis in several populations', *Psychometrika* 36(4): 409-426. doi:10.1007/bf02291366

Jöreskog, K.G. (1978) 'Structural analysis of covariance and correlation matrices', *Psychometrika* 43(4): 443-477. doi:10.1007/bf02293808

Jöreskog, K.G. and Sörbom, D. (2001) *LISREL 8: User's Reference Guide,* Lincolnwood: Scientific Software International.

Kaplan, D. (1990) 'Evaluating and modifying covariance structure models - a review and recommendation', *Multivariate Behavioral Research* 25(2): 137-155. doi:10.1207/s15327906mbr2502_1

King, G., Christopher J.L.M., Joshua A.S. and Tandon, A. (2004) 'Enhancing the validity and cross-cultural comparability of measurement in survey research', *American Political Science Review* 98(1): 191-207.

Little, T.D., Slegers, D.W. and Card, N.A. (2006) 'A non-arbitrary method of identifying and scaling latent variable in SEM and MACS models', *Structural Equation Modeling* 13(1): 59-72. doi:10.1207/s15328007sem1301_3

MacCallum, R.C., Browne, M.W. and Sugawara, H.M. (1996) 'Power analysis and determination of sample size for covariance structure modeling', *Psychological Methods* 1(2): 130-149. doi:10.1037/1082-989X.1.2.130

Marsh, H.W., Hau, K.T. and Grayson, D. (2005) Goodness of Fit in Structural Equation Models, in A. Maydeu-Olivares and J.J. McArdle (eds.) *Contemporary Psychometrics*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 275-340.

Marsh, H.W., Hau, K.T. and Wen, Z. (2004) 'In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings', *Structural Equation Modeling* 11(3): 320-341. doi:10.1207/s15328007sem1103_2

Marsh, H.W., Ludtke, O., Muthén, B.O., Asparouhov, T., Morin, A.J.S., Trautwein, U. and Nagengast, B. (2010) 'A new look at the Big Five factor structure through exploratory structural equation modeling', *Psychological Assessment* 22(3): 471-491. doi:10.1037/a0019227

Meade, A.W., Johnson, E.C. and Braddy, P.W. (2008) 'Power and sensitivity of alternative fit indices in tests of measurement invariance', *Journal of Applied Psychology* 93(3): 568-592. doi:10.1037/0021-9010.93.3.568

Meredith, W. (1993) 'Measurement invariance, factor analysis and factorial invariance', *Psychometrika* 58(4): 525-543. doi:10.1007/bf02294825

Meuleman, B. (2012) When are Intercept Differences Substantively Relevant in Measurement Invariance Testing?, in S. Salzborn, E. Davidov and J. Reinecke (eds.)

Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt, Heidelberg: Springer VS, pp. 97-104.

Millsap, R.E. (2011) *Statistical Approaches to Measurement Invariance*, New York: Routledge.

Millsap, R.E. and Everson, H.T. (1993) 'Methodology review: statistical approaches for assessing measurement bias', *Applied Psychological Measurement* 17(4): 297-334. doi:10.1177/014662169301700401

Muthén, B.O. and Asparouhov, T. (2013) *BSEM Measurement Invariance Analysis*. Mplus Web Note No. 17, available at http://www.statmodel.com/examples/webnote.shtml, accessed 11 January 2013.

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

Oberski, D.L. (2009) *Jrule for Mplus Version 0.91* (beta) [Computer software], available at https://github.com/daob/JruleMplus/wiki

Oberski, D.L. (2012) 'Comparability of Survey Measurements', in L. Gideon (ed.) *Handbook of Survey Methodology for the Social Sciences*, New York: Springer, 477-498.

Oberski, D.L. (2014) 'Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models', *Political Analysis* 22: 45-60. doi:10.1093/pan/mpt014

Piurko, Y., Schwartz, S. H. and Davidov, E. (2011) 'Basic personal values and the meaning of left-right political orientations in 20 countries.', *Political Psychology* 32: 537–561. doi:10.1111/j.1467-9221.2011.00828.x

Podsakoff, P.M., MacKenzie, S.B. and Podsakoff, N.P. (2012) 'Sources of method bias in social science research and recommendations on how to control for it', *Annual Review of Psychology* 63: 539-569. doi: 10.1146/annurev-psych-120710-100452

Révilla, M.A. and Saris, W.E. (2012) 'A comparison of the quality of questions in a face-to-face and a web survey', *International Journal of Public Opinion Research* 25(2): 242-253. doi:10.1093/ijpor/eds007

Saris, W.E. and Gallhofer, I.N. (2007) *Design, Evaluation, and Analysis of Survey Research,* Hoboken, NJ: John Wiley & Sons.

Saris, W.E. and Hagenaars, J.A. (1997) Mode Effects in the Standard Eurobarometer Questions, in W.E. Saris and M. Kaase (eds.) *Eurobarometer. Measurement Instruments for Opinions in Europe,* Mannheim: ZUMA, pp. 87-100.

Saris, W.E., Satorra, A. and Sörbom, D. (1987) 'The detection and correction of specification errors in structural equation models', *Sociological Methodology* 17: 105-129.

Saris, W.E., Satorra, A. and van der Veld, W.M. (2009) 'Testing structural equation models or detection of misspecifications?', *Structural Equation Modeling* 16(4): 561-582. doi:10.1080/10705510903203433

Schwartz, S.H. (1992) 'Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries', *Advances in Experimental Social Psychology* 25: 1-65. doi:10.1016/s0065-2601(08)60281-6

Schwartz, S. H., Caprara, G. V., Vecchione, M., Bain, P., Bianchi, G., Caprara, M. G., Cieciuch, J., Kirmanoglu, H., Baslevent, C., Lönnqvist, J-E., Mamali, C., Manzi, J., Pavlopoulos, V., Posnova, T., Schoen, H., Silvester, J., Tabernero, C., Torres, C., Verkasalo, M., Vondráková, E., Welzel, C. and Zaleski, Z. (2014) 'Basic personal values underlie and give coherence to political values: A cross national study in 15 countries.', *Political Behavior* 36(4): 899-930. doi: 10.1007/s11109-013-9255-z

Schwartz, S.H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O. and Konty, M.

(2012) ‘Refining the theory of basic individual values’, *Journal of Personality and Social Psychology* 103(4): 663-688. doi: 10.1037/a0029393

Schwartz, S.H., Melech, G., Lehmann, A., Burgess, S., Harris, M. and Owens, V. (2001) ‘Extending the cross-cultural validity of the theory of basic human values with a different method of measurement’, *Journal of Cross-Cultural Psychology* 32(5): 519-542. doi:10.1177/0022022101032005001

Schmitt, N. and Kuljanin, G. (2008) 'Measurement invariance: review of practice and implications', *Human Resource Management Review* 18: 210-222. doi:10.1016/j.hrmr.2008.03.003

Sörbom, D. (1989) ‘Model modification’, *Psychometrika* 54(3): 371-384. doi:10.1007/BF02294623

Steenkamp, J.-B.E.M. and Baumgartner, H. (1998) 'Assessing measurement invariance in cross-national consumer research', *Journal of Consumer Research* 25(1): 78-90. doi:10.1086/209528

Steinmetz, H. (2011) Estimation and Comparison of Latent Means across Cultures, in E. Davidov, P. Schmidt and J. Billiet (eds.) *Cross-Cultural Analysis: Methods and Applications*, New York: Routledge, pp. 85-116.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. and 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. doi:10.3389/fpsyg.2013.00770

Van de Vijver, F.J.R. and Poortinga, Y.H. (1997) 'Towards an integrated analysis of bias in cross-cultural assessment', *European Journal of Psychological Assessment* 13(1): 29-37. doi:10.1027/1015-5759.13.1.29

van der Veld, W.M. and Saris, W.E. (2011) Causes of Generalized Social Trust, in E. Davidov, P. Schmidt and J. Billiet (eds.) *Cross-Cultural Analysis: Methods and Applications*, New York: Routledge, pp. 207-247.

van der Veld, W.M., Saris, W.E. and Satorra, A. (2008) *JRule 2.0: User manual.* Unpublished document.

Vandenberg, R.J. (2002) 'Toward a further understanding of and improvement in measurement invariance methods and procedures', *Organizational Research Methods* 5(2): 139-158. doi:10.1177/1094428102005002001

Vandenberg, R.J. and Lance, C.E. (2000) 'A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research', *Organizational Research Methods* 3(1): 4-70. doi:10.1177/109442810031002

Vecchione, M., Caprara, G. V., Schwartz, S. H., Cieciuch, J., Schoen, H., Silvester, J., Bain, P., Bianchi, G., Kirmanoglu, H., Baslevent, C., Mamali, C., Manzi, J., Pavlopoulos, V., Posnova, T., Torres, C., Verkasalo, M., Lönnqvist J-E, Vondráková, E. and Alessandri, G. (2015) ‘Personal values and political activism: A cross-national study.’, *British Journal of Psychology* 106(1): 84-106. doi: 10.1111/bjop.12067

Table 1

Levels of Measurement Invariance Across Groups

Level of measurement invariance	The constraints it includes	The types of comparisons it allows if established
1. Configural invariance	The same latent variables are measured by the same items for all groups. No equality constraints are fixed.	No comparisons are allowed. It enables only an overall test of the similarity of the measurement models of the groups under study.
2. Metric invariance or weak measurement invariance	The factor loadings between the observed items and the latent variable are constrained to be equal across the compared groups.	Allows for the comparison of unstandardised relationships (unstandardised regression coefficients, covariances) between the constructs across groups.
3. Scalar invariance or strong measurement invariance	The factor loadings between the observed items and the latent variable and the indicator intercepts are constrained to be equal across the compared groups.	Allows meaningful comparisons of both unstandardised relationships and latent means across groups.

Table 2

Role of the Modification Index and the Power of the Test to Determine Parameter Misspecifications

	High power	Low power
Significant MI	Inspect EPC	Misspecification
Nonsignificant MI	No misspecification	Inconclusive

Note. Adopted from Saris et al. (2009); See also van der Veld and Saris (2011); MI = modification index

Table 3

Four Higher-Order Values, Ten Basic Values (Schwartz, 1992), and Sixteen Values with the PVQ-40 (Cieciuch et al., 2013)

4 higher-order values	10 basic values	Definitions of 10 basic values	16 values in the PVQ-40 with items to measure these values[12]
Openness to change	Self-Direction	Independent thought and action—choosing, creating, and exploring	Self-Direction—Action (sd11, sd34) Self-Direction—Thought (sd1, sd22)
	Stimulation	Excitement, novelty, and challenge in life	Stimulation (st6, st15, st30)
	Hedonism	Pleasure and sensuous gratification for oneself	Hedonism (he10, he26, he37)
Self-enhancement	Achievement	Personal success through demonstrating competence according to social standards	Ambition (ac24, ac32) Demonstrating success (ac4, ac13)
	Power	Social status and prestige, control or dominance over people and resources	Power (po2, po17, po39)
Conservation	Security	Safety, harmony, and stability of society, relationships, and self	Personal Security (se5, se21, se31) Societal Security (se14, se35)
	Conformity	The restraint of actions, inclinations, and impulses that are likely to upset or harm others and violate social expectations or norms	Conformity—Rules (co7, co28) Conformity—Interpersonal (co16, co36)
	Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provides	Tradition (tr20, tr25) Humility (tr9, tr38)
Self-transcendence	Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact	Benevolence (be12, be18, be27, be33)
	Universalism	Understanding, appreciation, tolerance, and protection for the welfare of *all* people and of nature	Universalism—Concern (un3, un8, un23, un29) Universalism—Nature (un19, un40)

Note. The item questions are available from the first author upon request.

Table 4

Global Fit Measures for the MGCFA

	χ^2	df	CFI	RMSEA	SRMR
Self-transcendence (BE-UNN-UNC)					
Configural	245.7	64	.948	.067 [.058-.076]	.041
Metric	254.2	71	.948	.064 [.056-.073]	.045
Scalar	270.4	78	.945	.063 [.055-.071]	.047
Openness (SDA-SDT-ST)					
Configural	111.3	22	.947	.080 [.066-.096]	.035
Metric	114.1	26	.948	.073 [.060-.087]	.038
Scalar	122.1	30	.946	.070 [.057-.083]	.041
Self-enhancement (HE-ACA-ACS-PO)					
Configural	208.4	58	.967	.064 [.055-.074]	.036
Metric	219.0	64	.966	.062 [.053-.071]	.040
Scalar	228.1	70	.965	.060 [.051-.069]	.042
Conservation (SES-SEP-COI-TR)					
Configural	172.1	42	.943	.070 [.059-.081]	.036
Metric	178.0	47	.942	.067 [.056-.077]	.038
Scalar	190.6	52	.939	.065 [.055-.075]	.040

Notes. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; for value abbreviations, see Table 3.

Table 5

Detection of Misspecification in the Invariance Test with Information Provided by the Mplus and Jrule Programs

		Mplus			Jrule			
					δ >.4		δ >.1	
Group	Parameter	MI	EPC	St. EPC	Power	Decision	Power	Decision
Detection of misspecification corresponding to the configural level								
On	UNN by be33	10.032	.193	.143	1.000	nm	-	-
On	UNC by be33	22.058	.454	.339	1.000	nm	-	-
On	UNN by un23	15.873	.230	.191	1.000	nm	-	-
P&P	HE by po2	12.210	.173	.144	1.000	nm	-	-
On	ACA by he26	20.101	.295	.175	1.000	nm	-	-
On	ACS by he26	11.144	.190	.123	1.000	nm	-	-
On	PO by he26	13.342	.235	.140	1.000	nm	-	-
On	SES by sc21	11.808	-.263	-.195	1.000	nm	-	-
Detection of misspecification corresponding to the metric level								
No modification indices								
Detection of misspecification corresponding to the scalar level								
On	se14	10.877	-.052	-.038	-	-	1.000	nm
On	se35	10.874	.063	.046	-	-	.999	nm
P&P	se14	10.880	.196	.144	-	-	.391	**m**
P&P	se35	10.876	-.173	-.126	-	-	.479	**m**

Notes. P&P – paper-and-pencil; On – online condition; MI – modification index; EPC - expected parameter change; St. EPC – standardised expected parameter change; δ (delta) – size of misspecification; Mplus – information provided by the Mplus program; Jrule – information provided by the Jrule program; UNC – universalism-concern; UNN –

universalism-nature; BE – benevolence; **m** – misspecification; nm – no misspecification; for item and value abbreviations, see Table 3.

Figure captions

Figure 1. A model testing for measurement invariance of two latent variables measured each by three indicators in two groups with the variances of the latent variables fixed to 1.

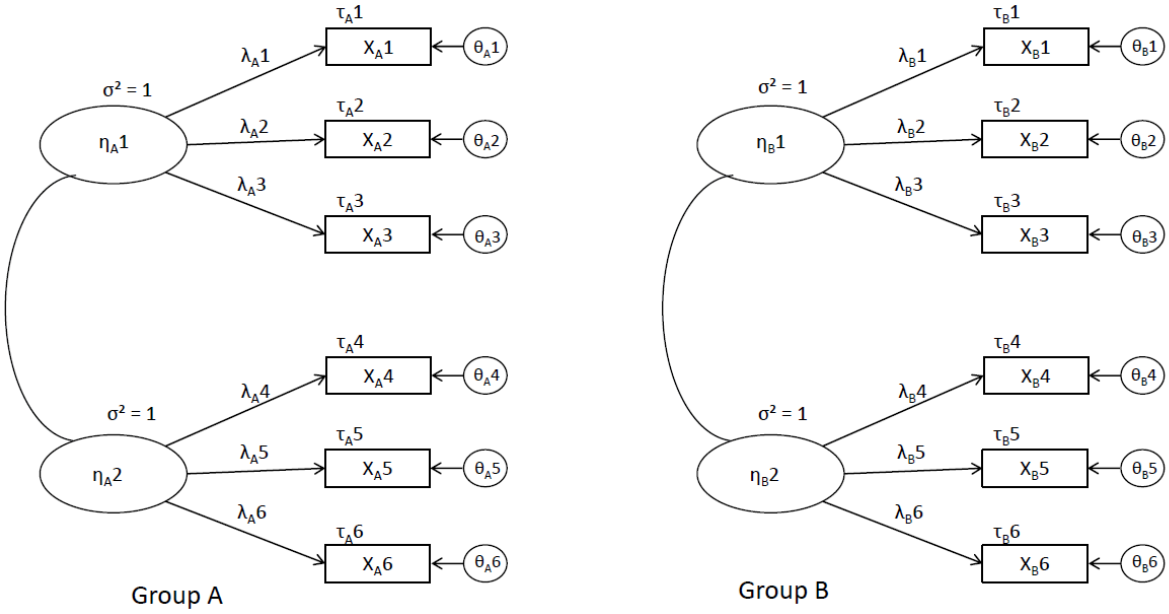
τA1
λA1
XA1
θA1
σ² = 1
ηA1
τA2
λA2
XA2
θA2
λA3
τA3
XA3
θA3
τA4
λA4
XA4
θA4
σ² = 1
ηA2
τA5
λA5
XA5
θA5
λA6
τA6
XA6
θA6
Group A
τB1
λB1
XB1
θB1
σ² = 1
ηB1
τB2
λB2
XB2
θB2
λB3
τB3
XB3
θB3
τB4
λB4
XB4
θB4
σ² = 1
ηB2
τB5
λB5
XB5
θB5
λB6
τB6
XB6
θB6
Group B

Figure 1

End-notes:

[1] https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5055&db=e&doi=10.4232/1.10202

[2] Partial invariance is supported when the parameters of at least two indicators per construct (i.e., loadings for partial metric invariance and loadings plus intercepts for partial scalar invariance) are equal across groups. According to Byrne et al. (1989) and Steenkamp and Baumgartner (1998), partial invariance is sufficient for meaningful cross-group comparison (but for different views see, e.g., de Beuckelaer and Swinnen, 2011; Steinmetz, 2011).

[3] For alternative, stricter criteria, see Meade et al. (2008). Recently, Muthén and Asparouhov (2013) and Asparouhov and Muthén (2013) proposed to use Bayesian and alignment methods to test for measurement invariance and partly rely on different global fit measures. However, we do not discuss these methods here and the reader is referred to the aforementioned web notes and to van de Schoot and colleagues (2013) as well as to Cieciuch et al. (2014).

[4] A Jrule version that adopts the output of the Lisrel program (Jöreskog and Sörbom, 2001) was developed by William van der Veld (available upon request from William van der Veld, email: W.vanderVeld@socsci.ru.nl); a Jrule version that adopts the output of the Mplus program (Muthèn and Muthèn, 1998-2012) was developed by Oberski (2009).

[5] Clearer guidelines as to which deviations may be tolerated should rely on future simulation studies. Recently, Oberski (2014) provided several guidelines to evaluate the sensitivity of parameters of interest to measurement (non)invariance (see also Meuleman, 2012, for a method to evaluate the sensitivity of latent means to scalar noninvariance).

[6] However, see, for example, Cohen (1988, 1992) who suggests a cut-off of 0.8.

[7] To the best of our knowledge, the literature is not clear about how many cross-loadings may be tolerated. However, the inclusion of cross-loadings for some of the groups is a threat to the assumption that the measurement operates similarly across groups.

[8] Further details about the sample and data collection may be obtained from the first author upon request. The data is available from the first author upon request.

[9] Fixing the variance of the latent variable to 1 or one of the items loadings as a reference to identify each latent variable did not have any impact on the results of the invariance test.

[10] Detailed output is available from the first author upon request.

[11] Indeed, only two items served as indicators for this value, so we could not rely on partial scalar invariance.

[12] The PVQ-40 scale that we used allows distinguishing between 16 values. The formulations of the items that measure these values can be obtained from the first author upon request.