The Index of Emancipative Values: Measurement Model Misspecifications

Supplementary Materials (SM)

This document contains Supplementary materials to the manuscript "The Index of Emancipative Values: Measurement Model Misspecifications." It comprises five sections (appendices). Appendix A reports the results of confirmatory factor analysis of the Index of Emancipative Values (EVI) using the data from the World Values Surveys. Appendix B describes the results of measurement invariance tests (a) of the EVI and its particular components across ten Welzel's cultural zones and (b) of pro-choice values across all countries participated in the WVS waves 1 to 6. Appendix C reports a simple simulated example illustrating how country-level measurement bias contributes to the strength of the aggregate-level correlations between attitudinal variables. Appendix D describes a re-analysis of Inglehart, Puranen and Welzel's (2015) finding regarding the association between country-average pro-choice values and country-average willingness to fight for one's country. Appendix E reports another simulation experiment which clearly shows that a theory-driven formative combination of several distinct constructs into a single higher-order construct may yield a measure which has high internal and external validity and fits theoretical predictions, but miss some important aspects of the reality at the same time. All references, figures and tables cited in these four appendices are found in the end of Supplementary materials. Feedback to my analysis is highly welcome; if you wish to provide your feedback, please write to bssokolov@gmail.com.

Contents

Appendix A: Confirmatory Factor Analysis of the EVI.

Welzel does not justify his index based on its internal consistence. Still, he reports the results of a hierarchical (second-order) exploratory factor analysis of twelve items defining the EVI using the country-pooled individual level data of 95 countries surveyed at least once by the WVS/EVS, based on the latest available survey from each country (from the period of 1995–2005). Factor loadings produced by that procedure are reported in Welzel (2013, 71, Table 2.3) and are quite high both at the first level and at the second level (all > 0.5 and almost all > 0.7). Below I replicate his analysis using the data of all 99 countries surveyed during the 3rd to 6th rounds of the World Values Survey (1995–2014). Unlike Welzel, I use all available surveys for each country, rather than only the latest one. Thus, my sample is larger than the one used by Welzel to validate the EVI, and covers a wider period. According to modernization-emancipation theory, values should not change significantly over short periods of time (Inglehart and Welzel 2005; Alemán and Woods 2016), and the use of a larger dataset provides additional opportunities for validation of the index, as well as for the assessment of the index's performance at different points in time.

I conduct a set of second-order confirmatory factor analyses (CFA) with the pooled sample for the 3rd-6th WVS rounds and for each particular round separately.[1] I use CFA because this procedure has some important advantages over the method of exploratory hierarchical factor analysis used by Welzel. First and foremost, CFA, being primarily a theory-driven analytic procedure, (a) directly assesses how a given a priori specified theoretical measurement model

[1] It should be noted that both within the total sample and within each wave, respondents are clustered within countries, and therefore multilevel CFA seems to be a more relevant method. I, however, use simple one-level CFA, mainly because Welzel, and also Aléman and Woods (2016), use this approach. Another reason is the very high computational complexity of multilevel CFA in large samples, not compatible with my hardware. It is worth noting that, being a simpler method than multilevel CFA, one-level CFA is therefore a more tolerant approach in terms of goodness of fit. As such, the pooled data results showing lack of fit are unlikely to be challenged by multilevel analysis.

fits the data, (b) uses (relatively) straightforward criteria of model rejection (such as RMSEA[2] and CFI[3]), and (c) allows for direct comparison between models of different complexities, while exploratory procedures seek only to find a factor solution that maximizes the amount of explained variance in observed indicators, not necessarily correlated with each other. In addition, CFA has several technical advantages: 1) it allows each indicator to have its own error variance; 2) it allows testing for the presence of non-zero cross-loadings of indicators on different components; 3) the multi-group CFA approach (MGCFA), an extension of basic CFA, allows for easy estimation and comparison of models for different subgroups of the total sample (Brown 2006; Alemán and Woods 2016).

The results of CFA for the pooled and round-specific samples are given in Table A1. In respective models, many standardized factor loadings are lower than 0.5 (e.g., for the first-order factor "Voice", all but one loading is lower than 0.5), and some are lower than 0.3.[4] Within-factor differences in standardized loadings across waves may be as high as 0.49 (for "equality"), and a between-wave range may be as high as 0.26 (second-order loading for "voice"). Finally, patterns of high- and low-loaded indicators do not completely coincide in various WVS rounds.

All models fit well according to RMSEA. The highest RMSEA value is 0.049 for the 4th round; all other models have RMSEA significantly smaller than 0.05. Moreover, the upper confidence bound for RMSEA is also lower than 0.05 for all but one model. However, the sample size is very large (more than 300,000 for the pooled-sample model and more than

[2] RMSEA stands for "Root Mean Square Error of Approximation". This measure varies from 0 to 1, with values lower than 0.05 indicating an acceptable model fit (Browne and Cudeck 1993, 144).

[3] CFI stands for "Comparative Fit Index". This measure is not restricted to a 0 to 1 range, but CFI values closer to 1 (typically, higher than 0.9 or 0.95) indicate a good fit (Hu and Bentler 1999).

[4] In various methodological papers, the value of 0.30 is referred to as a minimal (i.e., most tolerant) cut-off for meaningful factor loadings (Comrey and Lee 1992; Tabachnick and Fidell 2007; Brown 2006)

50,000 for every round-specific model), and it is known that RMSEA decreases as sample size increases. For example, Chen et al. (2008) showed in a simulation study that, when RMSEA < 0.05 is used as a cutoff, rejection rates converge to zero at sample sizes of 800 and above even for highly misspecified models (Chen et al. 2008, 476). So for sample sizes of 50,000 and above, as in the present study, the use of the conventional threshold of 0.05 may be meaningless. By contrast, according to CFI no model reaches the benchmark of 0.95, and only two models have CFI greater than 0.9 (these two are based on the pooled data and the 5th round's data).

Overall, the goodness-of-fit measures from pooled CFAs show that Welzel's original measurement model for the EVI is not decidedly unacceptable, but still obviously misspecified in some way. A particularly helpful tool that enables detection of the misspecifications of a CFA model can be found in the so-called modification indices, or MIs. MIs show the improvement in model fit (decrease in the model's chi-squared value) if a particular coefficient fixed to zero were to become unconstrained. Because the MI can be conceptualized as a chi-squared statistic with one degree of freedom, indices of 3.84 or greater (which reflects the critical value of the chi-squared distribution with one degree of freedom at $p < .05$) suggest that the overall fit of the model could be significantly improved if the fixed or constrained parameter was freely estimated (Brown 2006, 122). MIs, however, are sensitive to sample size, so for large samples higher cut-off values are recommended, such as 5 or 10.

For the model based on the data pooled across all waves, 135 MIs exceed 10, and 111 MIs exceed 100, indicating that the inclusion of multiple additional cross-loadings, residual covariances and direct effects of the second-order factor on the manifest variables will improve the model fit significantly. Such a large number of additional parameters to be freed according to MIs obviously indicates various types of model misspecification.

Thus, a large number of non-zero cross-loadings means that several items do not discriminate well between the first-order factors and may either be related directly to the second-order construct or simply not be useful for measuring that construct. In addition, a large number of non-zero residual covariances may indicate multidimensionality of the latent trait, or the fact that some model with alternatively defined first-order factors fits data better (Asparouhov, Muthén, and Morin 2015).

Appendix B: Tests for Invariance of Emancipative Values

B1. *What Is Measurement (non-) Invariance?*

As Davidov et al. (2014, 58) define it, "Measurement invariance is a property of a measurement instrument (in the case of survey research: a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups of respondents". More formally, established measurement invariance ensures that individuals from different groups that have the same score on a latent scale will provide similar responses on observed indicators, and *vice versa*, that those who have different scores on a latent variable will give consistently different responses. Consider a standard MGCFA model which is given by

$$y_{ijg} = \upsilon_{jg} + \lambda_{jg}\eta_{ig} + \varepsilon_{ijg} \quad (\mathbf{1})$$

where y_{ijg} represents the response of the individual i from the group g on the item j, $\upsilon_{jg}$is the intercept for the item j in the group g, $\lambda_{jg}$is the factor loading for the item j in the group g, η_{ig} is the individual score on the latent variable η in the group g, and e_{ijg} represents the residual for the individual i and the item j in the group g.

Three ordered levels of invariance are most frequently used in MGCFA. "Configural" invariance is the first and lowest level. It requires only that the loading patterns are the same across groups (that is, the same indicators have non-zero loadings on the same constructs in all groups). The second level of invariance is called "metric" or "weak" invariance. It requires that factor loadings are equal across groups, that is $\lambda_{jg} = \lambda_{jg'}, g \neq g'$ for all j and g. Finally, the third level of measurement invariance —"scalar", or "strong" invariance—assumes that not only loadings, but also the indicator intercepts are equal across groups, that is $\lambda_{jg} =$

$\lambda_{jg'}$and $\nu_{jg} = \nu_{jg'},\ g \neq g'$for all j and g (Steenkamp and Baumgartner 1998).[5]

In short, configural invariance ensures that a proposed model measures the same construct in all groups. Metric invariance ensures the cross-group equality of the intervals of the scale on which the latent variable is measured. It implies that an increase of one unit on the measurement scale has the same meaning in all groups (Davidov et al. 2014, 63). Scalar invariance ensures additionally that the origins of the latent scales are the same in all groups or, to put it another way, that group differences in latent means consistently manifest themselves in group differences in the means of the observed items (Steenkamp and Baumgartner 1998, 80). While other types of invariance can be assumed and tested [e.g. invariance of residual variances $\sigma_{jg}(\varepsilon_{ijg})$ across groups], it is generally considered that establishing full scalar invariance is sufficient to guarantee the reliability of latent means comparison across groups.

If metric invariance does not hold it implies the scale of the latent variable is different across groups. The left panel of Figure B1 gives a graphical illustration to the concept of metric non-invariance[6]. It displays the regression lines relating the scores on the observed item y_1 to the scores on the latent variable η in two groups. The factor loading (regression slope) in Group 1 is higher than the factor loading in Group 2, and the intercept in Group 1 is higher than the intercept in Group 2. The consequences of these group differences in intercept and factor loading are evident. First, predicted scores on y_1 for individuals in Group 1 with a certain

[5] It must be noted that, because the EVI is defined, among others, by nine essentially categorical indicators (those used to measure first-order constructs "autonomy", "equality", and "choice"), the measurement model for that index involves one more type of model parameter for which one should establish invariance, the variable thresholds (Millsap and Yun-Tein 2004). In addition, the EVI is a second-order construct. Thus, one should first establish invariance of the first-order factors, and then proceed with testing for invariance of the second-order factor. These details, however, are not so important since, as Tables B1-B4 below illustrate, the weakest assumption of configural invariance does not hold for the EVI, making tests for more demanding levels of invariance unnecessary.

[6] The example is borrowed from Wicherts and Dolan (2010).

score on η are always higher than predicted scores on y_1 for individuals in Group 2 with the same score on η. Furthermore, the higher score on η the higher difference in the predicted scores on y_1 between the individuals in Group 1 and individuals in Group 2 with the same latent score. *Vice versa*, a certain observed score on y_1 in Group 1 always corresponds to a higher score on η than the same score in Group 2, and the higher shore on y_1 the higher difference in corresponding latent scores between groups. Thus, the difference in the observed individual scores is not due to true differences in the score on the latent variable η but is produced by the interaction of group differences in the origins (intercepts) and in the intervals (loadings) of the scale of the latent variable.

Figure B1 about here

The right panel of Figure B1 gives an illustration of scalar non-invariance. Now factor loadings are the same in both groups, but the intercept in Group 1 is still higher than the intercept in Group 2. Thus, predicted scores on y_1 for individuals in Group 1 with a certain score on η are always higher than predicted scores on y_1 for individuals in Group 2 with the same score on η, and the amount of bias is equal to the difference between group-specific intercepts, ν_{11} and ν_{12} for all possible values of η. So even if metric invariance holds but scalar invariance does not, it implies that the latent mean scores from different groups are still not directly comparable since they are systematically upward or downward biased. The bias can arise from group differences in response style or various *method* factors[7] (Stegmueller 2011; Van Vlimmeren, Moors, and Gelissen 2016) or reflect specific cultural influence on understanding of the construct among the representatives of the group (Davidov et al. 2014).

Only if both factor loadings and intercepts are equal across groups or, in other words, the

[7]Here the term "method factor" refers to different aspects of survey design/conduction, such as translation, interview mode, sampling procedure, response rate, etc., that may uniformly influence individual responses in some countries and thus systematically bias true individual preferences in those countries (Stegmueller 2011; Brown 2015).

scale of the latent variable has the same unit of measurement and origin in each group, latent means can be meaningfully compared across groups. Unfortunately, in practice it is often impossible to establish full metric and especially full scalar invariance (Davidov et al. 2012; van de Shoot et al. 2013).

Although some authors (e.g. Meuleman 2012; Oberski 2014) argue that the amount of bias due to non-invariance in many practical instances is not critical (that is, biased measurement does not always lead to wrong substantive conclusions), most methodologists recommend to test for so called *partial invariance* in cases when it is impossible to achieve full invariance (Steenkamp and Baumgartner 1998; Vandenberg and Lance 2000). The concept of partial invariance was introduced in Byrne, Shavelson, and Muthén (1989) as "a compromise between full measurement invariance and complete lack of measurement invariance" (Steenkamp and Baumgartner 1998, 81). According to those authors, group-specific latent means can be validly compared when at least two items per construct function invariantly, that is have identical loadings across groups (including that fixed at unity for identification; this particular situation is referred to as *partial metric invariance*) and in addition have identical intercepts across groups (this situation is referred to as *partial scalar invariance*). Davidov et al. (2014, 66), however, note that "a few studies have indicated that partial equivalence may not be sufficient for meaningful cross-group comparisons… Further simulations are needed to provide a more informative recommendation for how applied researchers should handle partial measurement equivalence when full equivalence is not given."

B1.2. *Approximate (Bayesian) measurement invariance*

A recently proposed promising alternative to partial invariance testing in situations when full invariance is not supported by the data is a so called approximate Bayesian approach (Muthen and Asparouhov 2013; van de Shoot et al. 2013). In Bayesian statistics, model parameters are assumed to be probability distributions. These distributions (also called posterior

distributions) are products of two components, the prior distribution and the likelihood, or evidence from the data (Gelman et al. 2013). Prior distributions quantify a researcher's initial subjective uncertainty or degree of belief about the true parameter values, then updated by sample data to form posterior conclusions (Western 1999). Priors can be non-informative or informative. Non-informative priors reflect a high level of uncertainty about parameter values. They do not favor any specific area of parameter space, thus allowing the data component to dominate in final estimates. A uniform distribution and a normal distribution with zero mean and very large variance (e.g. 10^6) represent two probably most popular non-informative prior distributions in social science applications of Bayesian methods. When non-informative priors are used, the posterior parameter estimates are almost identical to those obtained with the frequentist approach (e.g., maximum likelihood estimates; see Figure B2, left panel). When informative priors are applied, the posterior estimates represent a trade-off between the researcher's prior belief in what the true parameter values have to be equal to and evidence from the data: the more informative prior the closer the posterior estimates to the *a priori* defined values and farther from those observed in the data (see Figure B2, right panel).

Figure B2 about here

Approximate invariance testing exploits this key distinct feature of Bayesian approach by imposing informative prior distributions on a specific family of auxiliary MGCFA model parameters: differences between substantive parameters (i.e., factor loadings and intercepts) across groups. Roughly speaking, the classical frequentist approach to invariance testing may be re-formulated in Bayesian terms as involving a very strong prior assumption that all differences in loadings and intercepts across groups are exactly zeros (see Figure B3, left panel). When the classical approach fails to establish invariance, a researcher can relax this strong assumption and allow a small variation in parameter values across groups by specifying a prior distribution of between-group differences with zero mean and relatively

small variance (Figure B3, right panel). For example, prior variance of 0.01 suggests that 95% of all group-specific deviations for a certain model parameter fall in the interval $[-1.96 * \sqrt{0.01}, 1.96 * \sqrt{0.01}] = [-0.196, 0.196]$ in unstandardized values, which is not so large to inevitably cause substantial bias in the estimates of latent means (van de Shoot et al. 2013).

Figure B3 about here

If a model with some acceptable level of invariance (defined by prior variance on parameter differences) fits sufficiently well, one can conclude that approximate invariance holds in the data. Both simulation studies and empirical research suggest that the prior variance on between-group parameter differences as high as 0.05 does not introduce substantial bias in the estimates of latent means and therefore allows for meaningful comparisons across groups (Muthen and Asparouhov 2013; van de Shoot et al. 2013; Davidov et al. 2015; Cieciuch et al. 2014; Zercher et al. 2015). The approximate Bayesian approach to invariance testing has been shown in a few recent papers to be quite flexible in handling relatively small (but numerous) between-group differences in model parameters. For example, using this method, Cieciuch et al. (2014; 2017) manage to establish cross-national approximate measurement invariance of the 19 basic human values (Schwartz et al. 2012); Davidov et al. (2015) demonstrate the comparability of attitudes toward immigration in the European Social Survey; and Zercher et al. (2015) show the comparability of the universalism value (Schwartz et al. 2012) over time and across countries in the European Social Survey.

B1.3. *Empirical criteria for deciding about (the lack of) measurement invariance*

In the context of the classical, or exact, approach to invariance testing, two main approaches for assessing whether measurement invariance holds in the data have been advanced in the literature. The first approach utilizes the fact that a model assuming a stronger form of invariance is nested within a model assuming a weaker form of invariance. It relies on the chi-

square difference test to determine do additional equality constraints required by the assumptions of metric and scalar invariance affect model fit. Statistically significant chi-square differences suggest that a model imposing less equality constraints fits data better than a presumably invariant model, therefore indicating lack of invariance. Chi-square difference test, however, is criticized by various authors (Cheung and Rensvold, 2002; Davidov et al., 2014) because it tends to overestimate the discrepancy in goodness-of-fit between nested models in large samples, which are the common case in comparative survey research.

An alternative approach is to use differences in global fit indices, such as CFI or RMSEA, between models assuming different levels of invariance. According to Chen (2007), if the sample size is larger than 300, metric non-invariance is indicated by a change in CFI larger than .01, when supplemented by a change in the RMSEA larger than .015 compared with the configural equivalence model. With regard to scalar invariance, non-invariance is evidenced by a change in CFI larger than .01 when supplemented by a change in RMSEA larger than .015 compared with the metric invariance model (see also Davidov et al. 2015, 250).

In Bayesian MGCFA, model fit can be evaluated using (1) the posterior predictive p-value (PPP)[8] and (2) the credibility interval (CI) for the difference between the observed and the replicated Chi-square values (Muthen and Asparouhov 2013; van de Shoot et al. 2013). Poor global fit indicates that actual parameter differences are larger than those that the researcher allows in the prior distribution, thus suggesting non-invariance (Davidov et al., 2015). In the context of Bayesian structural equation modeling, it is recommended that the non-significant

[8] In the context of Bayesian structural equation modeling, posterior predictive p-value represents a proportion of MCMC iterations for which the following inequality holds: $f(Y, X, \pi_i) < f(Y_i^*, X, \pi_i)$, where $f(*)$ is a fit statistic given by the standard likelihood-ratio chi-square test of an H_0 model against an unrestricted H_1 model, Y represents the data, X represents covariates that are conditioned on in the analysis, π_i represents the estimated parameter values at iteration i, and Y_i^* represents a new data set of the same size as the original data generated using parameter values at iteration i (Muthen and Asparouhov, 2012, 315). For a more general discussion of posterior predictive checking see Gelman et al. (2013).

PPP (that is, higher than zero, or, relying on a more conservative cut-off value, higher than 0.05), supplementing by the 95% CI which contains zero, indicates an acceptable model fit. An excellent-fitting model is expected to have a PPP around 0.5 and zero falling close to the middle of the CI (Muthen and Asparouhov 2012).

B2. *Tests for Cross-Zone Invariance of the EVI and Its Components*

I follow the empirical strategy of Aléman and Woods but extend the spatial coverage of their analysis. I test the invariance of the EVI across all ten of the cultural zones defined by Welzel. These zones are the Islamic East, the Indic East, the Sinic East, the Orthodox East, the Old West, the Reformed West, the Returned West, the New West, Latin America and sub-Saharan Africa.[9] Aléman and Woods present the results of MGCFA for emancipative values only for four zones.[10] While most comparative analyses involving emancipative values are executed at the national level, focusing on cross-zone invariance is adequate for two reasons.

First, in their theories of cultural change, Inglehart and Welzel place specific attention on supra-national cultural entities, reflected in their famous cultural map of the world (Inglehart and Welzel 2005, Ch. 2), of which Welzel's concept of cultural zones is an updated version. They base the classification of cultural regions upon a scatter plot of country mean scores on two value dimensions, one of which is emancipative values (in the latest, Welzel's version; the other cultural dimension is secular values). A test for cross-zone invariance of the EVI can serve as a validity check for the proposed classification of cultures. In particular, if CFA models for different zones are not equivalent, then the observed zone-specific mean scores on

[9] For a detailed list of the countries belonging to each zone see Welzel (2013, Table 1.3).

[10] These zones are the New West, the Old West, the Reformed West, and Sub-Saharan Africa. Aléman and Woods also did not report certain important details about their analytical procedures. Thus, it is not clear from their paper whether they accounted for the categorical nature of the indicators used to define emancipative values. Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, I use the WLSMV estimator for parameter estimation, which is the default option for dealing with categorical or non-normal responses in many structural equation modeling software packages (including MPLUS 7.11, which I use).

emancipative values, defining what Welzel calls the "gravity center" of each zone, are not generally comparable and likely biased (in an unknown direction and to an unknown, but potentially large extent) along the Y-axis of the cultural map presented in Welzel (2013, 89).[11]

Second, Aléman and Woods report that they failed to achieve model convergence when trying to estimate the MGCFA model for all 98 WVS countries simultaneously. Non-convergence is a common problem in MGCFA when the number of groups is large. In such cases, country-by-country CFAs can be used to examine the coherence of value patterns across different societies (Davidov et al. 2014, 65). However, presenting and comparing the results of country-by-country analyses even for a subsample of sixty countries participating in the 6th wave of the WVS is time- and space-expensive, and may also raise interpretation problems for readers. In addition, parameter estimates from country-specific analyses can be compared across societies only visually, which is not problematic when configural invariance is being assessed, but prevents a researcher from conducting formal tests for metric/scalar invariance.

Fortunately, WVS countries are clustered within cultural zones not randomly, but following a theoretical expectation that national value patterns will be more coherent within than between zones due to shared historical legacies (Welzel 2013, 120–139). Hence, finding cross-zone invariance would imply that the minimally acceptable level of comparability for emancipative values holds and therefore a cumbersome within-zone analysis might be worthwhile for obtaining more nuanced evidence. Conversely, finding cross-zone non-invariance for the EVI would ultimately imply non-invariance at the national level, where even more potential sources of non-equivalence are introduced. So testing for cross-zone invariance is simply a more parsimonious way of achieving the same goal as country-level analysis. For ease of

[11] In this paper I study the measurement validity of emancipative values, since that concept is used a little more frequently in applied political research. However, as Aléman and Woods' findings suggest, secular values are also problematic from the viewpoint of measurement validity. As such, my analysis here reflects a more general problem with Inglehart and Welzel's cultural map.

exposition, I discuss here only the results for the 6th wave of the WVS (which are presented in Table B1 in the end of Supplementary materials). However, my main conclusions hold for other rounds as well (see Tables B2-B4).

CFA models for the EVI, estimated separately for each specific cultural zone, show that even the weakest assumption of configural invariance does not hold for the index across zones (Table B1). The model including all twelve indicators, with all loadings high enough (and having the theoretically expected direction), does not fit well in any cultural zone. First, I have to exclude several items in certain zones to achieve model convergence. In each zone, irrespective of whether a full or reduced set of items is used to define the index, at least one item (usually two or three) has a factor loading lower than 0.30. In some zones there are several non-significant or even negative first-order loadings.

In some other zones, several second-order loadings are either non-significant or lower than 0.30. Among the observed items, the most problematic are "when jobs are scarce, men should have priority over women to get a job" (excluded in three zones, the New West, the Reformed West, and the Returned West, in order to achieve model convergence, and with loadings lower than 0.30 in two other zones), and the item which measured the priority of "protecting freedom of speech" (excluded in five zones in order to achieve model convergence, and with loadings lower than 0.30 in two other zones).

Among the first-order factors, "autonomy" and "voice" seem to be the most problematic, which is to be expected. Both factors are defined by items measured with ranking, and conventional factor analytic procedures are in general not applicable to such scales (Jackson and Alwin 1980). This can partly explain the low loadings of the observed items on those factors, as well as the relatively low second-order loadings of "autonomy" and "voice" on emancipative values. However, non-invariance is not caused by low loadings; it is caused by

variation in loading sizes across groups.[12] In addition, for "equality", which is defined by conventional Likert-type items, configural invariance does not hold either. In general, only one component of emancipative values, "choice", has high-loaded indicators across all cultural zones. I then test for metric and scalar invariance of this sub-index.

The classical approach suggests that neither full metric nor full scalar invariance hold across ten cultural zones for "choice": the RMSEA and CFI of the configural model are 0.000 and 1.000 respectively; for the metric model they are 0.126 and 0.911 respectively (Δ RMSEA = 0.126; ΔCFI = 0.089), and for the scalar model they are 0.199 and 0.557 respectively (Δ RMSEA = 0.073; Δ CFI = 0.354). The chi-square differences between all these models are also significant.

In contrast, the Bayesian approach suggests that "choice" is approximately metric and scalar invariant across ten cultural zones. The model assuming a normal prior with zero mean and 0.01 variance[13] for the differences in factor loadings and intercepts across ten cultural zones has the PPP = $0.113\ (> 0.05)$. The CI for the difference between the observed and the replicated chi-square values is $[-16.151;\ 66.954]$[14], that is, it includes zero. It is nevertheless worth noting that in some cultural zones several loadings and intercepts considerably deviate

[12]Ippel, Gellisen and Moors (2014) explored longitudinal and spatial invariance of the four-item post-materialism scale (which is the same as Welzel's "voice" index) using the proper Jackson-Alwin method for ipsative data, and nevertheless found lack of invariance for that scale across ten Western European countries.

[13] Notice that this is a more conservative (that is, assuming a stricter level of approximate invariance) threshold than 0.05 used in other similar studies.

[14]To identify the model I use the marker variable method. I constrain the factor loading of the item "Divorce" to one and its intercept to zero in all groups. The latent means and variances were freely estimated in all groups. To test the sensitivity of the results to the choice of the marker, I re-run the model two more times, each time with a different item as the marker item. When the item "Abortion" is used as the marker variable, the essential conclusion remains the same, but the PPP value and the chi-square difference CI change in slightly unfavorable direction: PPP = 0.031 and CI is $[-1.953; 85.795]$. When the item "Homosexuality" is used as the marker variable, PPP = 0.000 and CI does not include zero, that means that the choice of the marker matters. However, when a slightly higher level of prior variance (0.05 instead of 0.01) is imposed, all three models fit well, so the general conclusion supporting approximate invariance is reliable.

from the prior defined parameter values (see Table B5). Despite these deviations the PPP and CI indicate acceptable model fit, which suggests that both approximate metric and scalar measurement invariance hold for pro-choice values across WVS cultural zones[15].

B3. *Tests for Cross-National Invariance of "Choice"*

Approximate scalar invariance of "choice" across different cultural zones is an encouraging finding. Yet, as noted above, most applied analyses involving the notion of values are conducted at the nation level. I therefore perform six Bayesian MGCFAs, one for each separate wave of the World Values Surveys, to test for the cross-national comparability of "choice". Number of countries and respondents included in the analysis in each wave are reported in Table B8.[16]

Columns 2 and 3 in Table B6 show respectively the PPPs and the CIs for the difference between the observed and the replicated chi-square values for six wave-specific MGCFAs of the "choice" value dimension. All models assume zero mean and 0.05 prior variance for the differences in the item intercepts and loadings between countries. According to the fit measures, approximate scalar measurement invariance is supported for the WVS waves 1, 2, 3, and 5. The CI for the model for the Wave 6 contains zero, but the PPP for the same model is lower than 0.05. Finally, for the model for the Wave 4 both fit indices show unacceptable fit,

[15] Differences of size 0.3-0.4 (even in unstandardized values) between country-specific factor loadings or intercepts and their sample average value, which one can find in Table B5, may appear as tremendous ones. Nevertheless, varying (cross-nationally) contributions of particular indicators to the latent country means are accounted for in Bayesian MGCFA. In addition, one can simply remove the most outlying unit(-s) from the sample or try to establish partial (metric or scalar) approximate invariance by releasing approximate equality constraints for problematic items (that is, loadings and/or intercepts for which the highest country-specific deviations from the sample average parameter values are observed).

[16] I do not apply frequentist MGCFA to the same samples, as it was done [for the purpose of comparison between the two approaches] in similar applications of the Bayesian approximate invariance testing by Davidov et al. (2015), Cieciuch et al. (2014), and Zercher et al. (2015). I have already shown that the classical approach failed to establish scalar invariance of "choice" even across ten cultural zones, so it is quite reasonable to anticipate that it will fail at the country level (much more heterogeneous) as well.

which indicates non-invariance.

Table B6 about here

I then test for partial approximate invariance by releasing from prior constraints the loading and the intercept for the item reflecting how justifiable do people consider homosexuality. Columns 4 and 5 in Table B6 present the fit statistics for the round-specific MGCFA models assuming partial approximate invariance of pro-choice values. Now all models except that for the Wave 4 fit well, according both to the PPPs and the CIs. The CI for the model for the Wave 4 does not contain zero and the PPP for the same model is still lower than 0.05, which indicates relatively high differences in parameter values across the countries covered by that wave.

Because even partial approximate invariance is not supported for the WVS Wave 4, I use group-specific PPPs to identify countries which are clearly different from the rest of the sample. After excluding the four countries with the lowest group-specific PPPs (which are Algeria, Bangladesh, Pakistan, and Saudi Arabia) model fit of the respective round-specific model becomes acceptable. I apply the same procedure to the model for the Wave 6, which is also problematic in terms of goodness of fit. Again, after excluding the six countries with the lowest group-specific PPPs (which are Bahrain, Jordan, Lebanon, Morocco, Pakistan and Palestine), model fit improves considerably. Interestingly, all countries identified as the most dissimilar are Islamic states (see discussion below, in Section B4 of this Appendix).

Figure B4 about here

The correlations between the raw country mean scores on "choice" and the mean scores based on the Bayesian MGCFA in WVS Waves 3 to 6 are shown in Figure B4. All correlations are quite high, though not perfect. Wave-specific correlations vary from 0.905 to 0.935, except that for Wave 4, which is the most non-invariant wave. For that wave the correlation between

the raw mean scores and the Bayesian mean scores is 0.817 and the correlation between the raw mean scores and the Bayesian mean scores based on the partial approximately invariant MGCFA model is 0.818. These results imply that country mean rankings based on the raw scores and on the Bayesian MCGFA estimates, while quite similar, are not completely equivalent. In contrast, the wave-specific correlations between the mean scores obtained from the MGCFA model assuming full approximate invariance and the mean scores obtained from the MGCFA model assuming partial approximate invariance are all higher than 0.99, thus suggesting that both approaches provide identical country rankings.

In addition, scatterplots presented in Figure B4 indicate a slightly non-linear (sigma-shaped) relationship between the raw mean scores and the Bayesian latent means in the WVS waves 3 to 6, which implies that the raw mean scores may overestimate the mean level of pro-choice values in countries with either low or high scores on those values and underestimate the mean level of pro-choice values in countries with medium scores on those values.

B4. *Why some countries are different?*

It has been noted above that Islamic countries appear to be the most probable source of non-invariance in the WVS with regard to the "choice" value. Investigation of the country-specific PPPs, provided by MPLUS software as the indicators of how well the measurement model under test fits in particular countries, supports this surmise. For example, in the Wave 4, which is the most non-invariant wave, the countries with the lowest PPPs are Algeria, Bangladesh, Pakistan, and Saudi Arabia, and in the Wave 6, which is the second most non-invariant, such countries are Bahrain, Palestine, Jordan, Lebanon, Morocco, and Pakistan. Importantly, all these countries are Muslim countries, though, according to Welzel's classification, they belong to two different cultural zones, the Islamic East and the Indic East.

What makes these countries different? Table B7 shows group-specific loadings and intercepts

for the items measuring individual acceptance of homosexuality and abortion for aforementioned wave-country combinations. For the item measuring perception of homosexuality, the group-specific factor loadings in most of these combinations are much lower than the sample average for the respective wave, as well as the group-specific intercepts are (though some exceptions exist). Conversely, for the item measuring support for abortion, the group-specific factor loadings in most these countries are in general higher than the sample average, as well as the group-specific intercepts (again, exceptions exist). This means that the acceptance of homosexuality correlates much less well with the "choice" dimension in several Islamic countries compared to the rest of the sample, while the acceptance of abortion correlates slightly stronger. It seems that pro-choice values in these countries are formed by only two components, the attitudes toward abortion and divorce, while the attitude toward homosexuality appears to be, to some extent, an independent attitude.

Table B7 about here

The most obvious explanation for that is the impact of religion. According to the Sharia law, homosexual activity is a crime, and in many Muslim countries it is illegal. Moreover, in several countries it carries the death penalty (Adamczyk and Pitt 2009, 339). At the same time, abortion and divorce, while condemned in general, are nonetheless allowed under certain circumstances (Bowen 1997). So it is quite logical that the attitude towards homosexuality does not associate strongly with the attitudes toward abortion and divorce in Islamic countries. This finding is not surprising. As Alexander et al. (2016, 911) note, "sexual freedoms remain an especially contested domain of emancipation because conservative forces, most notably religion, concentrate their resistance here". These authors also show that pro-choice values are actually more prevalent in more secular countries. They however emphasize the overall impact religion has on support for pro-choice values. In practice, some sexual freedoms may face stronger religious opposition than others, depending on the socio-

cultural context, which in turn affects the measurement of the overall construct.

Does bias due to the culture-specific religion effects mean that Islamic countries may not be meaningfully compared to other WVS countries in terms of prevalence of pro-choice values? It is worth noting that particular features of survey design/conduction in those countries do not cause the bias[17]. Instead, it reflects a substantive cultural effect on support for a particular sexual freedom, homosexuality, inherently specific to Muslim societies.

Consider, for example, a respondent, living in a country with a Muslim majority, who personally believes that homosexuality can be justified, at least to the same extent as abortion and divorce. However, due to severe social pressure, she has a lot of reasons not to reveal explicitly her true opinion on that issue, so she will probably choose the answer "completely unjustifiable" responding to the respective survey item, yet may choose more tolerant options when responding about her support for abortion and divorce. Though her individual response to this item is obviously biased downward, it should not distort much the validity of the mean score on tolerance toward homosexuality for the country she lives in – as well as the validity of the scores on the attitudes towards abortion and divorce – because the average level of tolerance is indeed low in that country.

Similarly, specific covariances between items measuring "choice" in Islamic states, though unalike to those observed in other countries, do not necessarily represent the measurement-related artifact. Instead, they reflect substantive differences in how many people living in Islamic states (do not) perceive a particular sexual freedom as an integral part of a broader domain of reproductive freedoms, compared to other cultural zones. In addition, low correlation between acceptance of homosexuality and "choice" in general and therefore small contribution of the respective item to the overall country means on "choice" in some Islamic states is accounted for in the approximate Bayesian model (albeit it isn't when the raw mean

[17] At least, it seems that they are not *fully* responsible for it.

scores are used). So the mean scores on "choice" for these states provided by Bayesian MGCFA are meaningful and can be compared with the mean scores for other WVS countries.

5. Limitations

The analysis above is not free from limitations. First, it establishes the comparability of the "choice" value only across each WVS wave separately, not simultaneously across all possible WVS country-wave combinations. However, the WVS data on values are often used for longitudinal studies. This requires that not only cross-national but combined cross-national and cross-temporal invariance of the measurement instrument does hold.

Unfortunately, due to internal memory limitations, MPLUS 7.11, which is used for model estimation in this paper, is unable to estimate a MGCFA model including all possible WVS round-country combinations (226 in total) as separate groups. Some previous findings, however, indicate that longitudinal invariance within one country is easier to establish than spatial invariance across countries in the same time point. For instance, using Eurobarometer data, Ippel, Gellisen and Moors (2014) explored longitudinal and cross-national invariance of the four-item post-materialism measure across ten countries and twenty years (1976–1997). They discovered that longitudinal *invariance held* in almost all countries in their sample, with the single exception of Denmark. However, they found evidence of a lack of invariance *across* countries. This suggests that the within-wave cross-national comparability of "choice" observed for each WVS wave separately can be considered as a reasonable basis for assuming the combined cross-national and cross-temporal comparability of this construct.

Second, there is still a lack of evidence from simulation studies about how small should be the prior variance on differences in factor loadings/intercepts across groups to conclude with certainty that the approximate invariance does hold (Cieciuch et al., 2014). The existing results suggest that the prior variance of 0.05 is sufficient to avoid substantial bias in latent

mean comparisons (Van de Schoot et al., 2013), and few practical applications of the method, including one in the presented paper, rely on that recommended cut-off value (e.g., Zercher et al., 2015, Davidov et al., 2015). I tried to impose stricter levels of invariance for each of the six wave-specific models. I found that, in general, my conclusion about cross-national invariance of pro-choice values remained robust to the use of more narrow prior variances, such as 0.01 to 0.03 depending on the WVS wave (see Table B9)[18].

Third, all three items measuring pro-choice values are 1 to 10 scales, which are either highly skewed to 1 or multi-modally distributed (with peaks at 1, 5 and 10) in most WVS countries. Bayesian algorithms implemented in MPLUS 7.11 currently cannot handle non-normal or ordinal variables and treats them as normally distributed continuous (Cieciuch et al., 2014). Unfortunately, no simulation studies of the robustness of the approximate Bayesian invariance testing to the non-normal data have been published to date, so it is not clear whether and to what extent the non-normality of the observed pro-choice indicators affect model fit[19].

Fourth, some recent research argues that the PPP may be a flawed measure of the model fit when the informative priors are used in Bayesian structural equation modeling, and a more sophisticated measure, known as a mixed prior posterior predictive p-value (or PPPP) should be used instead (Hoijtink and Van de Schoot 2017). Unfortunately, as of now, only MPLUS Version 8 (unavailable to the author) implements that alternative fit measure. So testing whether the findings from this analysis remain robust when more sophisticated measures of model fit are applied is an important task for future research

Nevertheless, strong and cross-nationally robust correlations between the items measuring

[18] It should be noted that assuming stricter prior variance in general worsens model fit, though not critically. In order to test sensitivity of my results to the choice of prior variance, I also re-ran each wave-specific model with a less restrictive prior of 0.1. I found that all those "liberal" models fitted well.

[19] This might be less of a problem for Bayesian analysis, because asymptotic properties of estimators are obtained from MCMC sampling, and no prior asymptotic assumptions are necessary.

pro-choice values definitely suggest that people living in different countries have something in common when they think about various sexual freedoms and respond to respective WVS questions. Unfortunately, “choice” is the only invariant component of the EVI in the WVS data. All other index components, as well as the second-order construct itself, fail to satisfy even the weakest requirement of configural invariance. Strictly speaking, this means that the EVI, in its current version, does not measure the same latent value dimension(s) in different cultural zones and, consequently, in different countries. Therefore, it should not be used for cross-national comparisons and substantive quantitative research.

Appendix C: A Simple Simulation Illustrating the Impact of Method Effects on the Strength of the Aggregate-Level Correlations

In the main text, I argue that the strong aggregate-level associations of the particular components of the EVI with each other and with other variables of interest can be partially attributed to measurement error. To offer a simple example of how the country-level measurement bias contributes to the strength of the aggregate-level correlations between attitudinal variables, consider the following simulation experiment. Please note that here I use the same notation as in the main text.

Let η_1 and η_2 be two random samples of 100,000 observations from two weakly correlated standard normal distributions (let $\rho_{\eta_1,\eta_2} = 0.25$)[20]. If one then quasi-randomly assigns these observations to 100 groups (assigning the n-th 1000 observations to group n, such that the first 1000 observations are assigned to group 1, and the hundredth 1000 observations are assigned to group 100), one will find that the correlation between the vectors of group means on η_1 and η_2 (let us denote them as μ_{η_1} and μ_{η_2}) is the same as between η_1 and η_2 themselves (even if random individual-level error is added in the data generation process).

If one then generates two vectors (each of length 100) of highly correlated country-level method effects u_{y_1} and u_{y_2}, and adds them to the vectors μ_{η_1} and μ_{η_2}, one will find that the aggregate-level correlation between the resulting vectors μ_{y_1} and μ_{y_2} increases considerably

[20] In their CPS paper, Welzel and Inglehart (2016, 1071) note that "the "choice" and "voice" components of emancipative values correlate at *R* = .22 at the individual level. By contrast, aggregate measures of these components correlate at an *R* of .62 between countries" They use this observation to justify their statement that "Weak and variable inter-item convergence within countries is (a) the norm and (b) irrelevant for convergence patterns that exist at the aggregate level between countries." In my simulation example, I choose the value of the individual-level correlation which is similar to that reported by Welzel and Inglehart, in order to show that, under quite realistic conditions, the aggregate-level measurement error may fully account for the observed discrepancy between the individual-level and the aggregate-level correlations of the same variables.

compared to the individual-level correlation between η_1 and η_2. For example, assume that u_{y_1} and u_{y_2} are samples from a bivariate normal distribution with zero means and unit variances, divided by a factor of 25,[21] and $\rho_{u_{y1},u_{y2}} = 0.75$ (which may be a rather mild assumption). In such a case, adding u_{y_1} and u_{y_2} to μ_{η_1} and μ_{η_2} increases the strength of the aggregate-level association between the observed vectors of group means, μ_{y_1} and μ_{y_2}, to 0.55,[22] that is, it more than doubles it compared to the individual level.

[21] I divide both vectors of country-specific biases by 25 to make the absolute range of the biases smaller than the range of individual responses (generated from the same multivariate distribution) and the range of the group mean vectors μ_{η_1} and μ_{η_2} (since the sample means of both η_1 and η_2 are equal to zero and observations are assigned to groups randomly, the range of group means is considerably smaller than the range of individual responses). In the reported setup, the latter is on average four times higher than the range of generated biases. I experimented with different values of the division factor, but even if larger factors of 50 or even 100 were used, the increase in the strength of the observed aggregate-level correlation remained non-negligible (20% in the latter case). Importantly, in the real world most survey items are ordered polytomous scales and the magnitude of method bias for a single item can be as large as one third of that item's range (as in the example with different response styles presented in the main text).

[22] This is an average value over 10,000 simulations. The 95% confidence interval for this quantity is from 0.41 to 0.68. Interestingly, in the absence of the country-level bias the macro-level correlation $\rho_{\mu_{\eta_1},\mu_{\eta_2}}$ is on average equal to the strength of the individual-level correlation ρ_{η_1,η_2} (mean = 0.25, CI = [0.06; 0.43]).

Appendix D: The Relationship between Pro-Choice Values and the Average Willingness to Fight for One's Own Country

Inglehart et al. (2015) provide cross-sectional, longitudinal and multi-level evidence of the strong negative association between pro-choice values and mass public's willingness to fight in wars[23]. In particular, using the latest available WVS survey for 79 countries they show that the correlation between the country means on pro-choice values and the average willingness to fight is equal to – 0.47. In a bivariate regression pro-choice values explain 22% of the variance in mean willingness to fight. Moreover, when two specific clusters of countries (the Nordic countries and the former Axis powers[24]) are accounted for, the explained variance grows to 65%. Importantly, the effect of pro-choice values is robust to inclusion of various economic and political controls. Furthermore, when included in the model together with pro-choice values, all other predictors, except dummies for aforementioned two clusters of countries, lose significance, which leads Inglehart et al. to the conclusion that "choice" is an important mediating variable that absorbs the "strong pacifying effect" of improving living conditions on willingness to fight. This finding confirms their basic hypothesis that "ascending life opportunities diminish willingness to fight in wars through their tendency to produce a choice-oriented culture", in which life is seen as a source of opportunities and pleasures rather than a source of threats, and which is therefore highly intolerant to human

[23]The proportion of respondents, expressed in fractions of 1, saying they are willing to fight for their country when responding to the following WVS question:

> *Of course, we all hope that there will not be another war, but if it were to come to that, would you be willing to fight for your country? (The response options are 'yes' and 'no'.)*

[24] In the former Axis powers publics demonstrate extremely low level of willingness to fight for country, which is probably the legacy of defeat in the Second World War. In contrast, in the Nordic countries mean readiness to fight is quite high, given their exceptional level of sexual emancipation. Inglehart, Puranen and Welzel explain the peculiarity of the Nordic countries by the threat of an "empire that represents an opposite way of life and has repeatedly shown its territorial ambitions" – previously the Soviet Union and now Russia. "Exposure to this threat keeps the willingness to defend their countries' lifestyles stronger than one would otherwise expect from publics with such pronounced pro-choice values as the Nordic ones" (Inglehart et al., 2015: 420)

costs caused by wars (Inglehart et al., 2015: 420, 428).

These authors, however, use raw mean scores to measure prevalence of pro-choice values, and I have shown above that raw means may provide biased country rankings compared to MGCFA-based means. Is their finding robust to the adjustment in the measurement model for "choice" using the Bayesian approximate approach? I replicate their cross-sectional analysis using the data from 58 countries[25] covered by the 6 Wave of the WVS. I find that the correlation between the latent country means on pro-choice values obtained with the Bayesian approximate approach and the average willingness to fight is statistically significant and equal to -0.443, which is close to the estimate reported by Inglehart et al. The adjusted R-squared for a bivariate regression of willingness to fight on pro-choice values is equal to 0.182 (also quite close to the original figure from Inglehart et al.'s paper). The Wave 6 covers only one Nordic country, namely, Sweden, and two former Axis powers, Germany and Japan. Sweden and Japan are indicated as outliers by a formal test (see also Figure D1). Excluding them increases the explained variance to 21%. The growth in the explained variances is far not as impressive as in the original paper by Inglehart et al., but they have five Nordic countries and five former Axis powers in their data, so the effects of these two particular historical legacies are more tangible. Importantly, it should be noted that the correlation between the raw means on pro-choice values and the average willingness to fight is equal to -0.499, thus indicating that the use of raw means instead of [Bayesian] MGCFA estimates leads to the overestimation of the association's strength by approximately 13 percent.

[25] Kuwait and Egypt are removed due to the fact that one or more indicators of pro-choice values are missed in the national WVS questionnaires for these countries in the Wave 6.

Appendix E: Are Explanatory Power and a Convincing Theory Sufficient When Assessing the Quality of Formative Constructs?

As I claim in the main text, "A potential undesirable consequence of the use of complex value measures is that, despite their complexity, such indices may oversimplify, or blur, actual associations between particular value dimensions and their expected correlates." To gain an intuitive sense of this, consider the following simple simulated example.

Let there be three latent variables, each defined by three observed indicators. Let the sample size be 100,000, all means be set to zero, and all loadings and residual variances be set to 1 (see Panel A in Figure E1). The first latent variable has a strong positive effect on an observed outcome variable (standardized $\beta = 0.7$), the second factor has a moderate positive effect ($\beta = 0.3$), and the third factor has a moderate negative effect ($\beta = -0.3$). The latent variables also positively correlate with each other ($\rho = 0.6$).

Then imagine that a scholar observing the relatively high correlations between the factors assumes that those factors should be combined into a second-order latent construct which is hypothesized to be positively related to the outcome. The respective model (Panel B in Figure E1) fits the data well according to the most popular global fit indices (RMSEA = 0.046 (95% CI: 0.44 – 0.46); CFI = 0.977, TLI = 0.968). It supports the hypothesis of the positive effect of the second-order construct on the outcome (estimated $\beta = 0.799$, standardized $\beta = 0.564$, p-value = 0.000).

Now consider another researcher, who follows the formative approach. She simply combines individual scores on all nine observed indicators, because all of those indicators are conceptually related to some theoretically meaningful quality, and then uses the composite score to predict the outcome. Again, the regression of the outcome on the composite score (Panel C in Figure E1) favors the theory of "only one construct positively related to the

outcome" (estimated $\beta = 0.606$, standardized $\beta = 0.445$, p-value = 0.000).

It is worth noting that the true model is unknown to both researchers, but the fitted models (a) show a good approximation of the real data according to conventional statistical criteria and (b) associate strongly with its theoretically expected correlate. Those models, therefore, are correct in terms of both the formative approach and the reflective approach. However, they do not reflect the presence of the first-order factor that has a moderate negative effect on the outcome. Nevertheless, most readers of a potential paper reporting those two models would agree that the substantive inferences from those models are reliable according to the current methodological standards.

Generally speaking, inferences based either on approximately "good" reflective measurement models or on measurement models defined according to the formative approach may have relatively high internal and external validity, but either may miss some important aspects of the reality at the same time. The detection of misspecifications of such models is not an easy task,[26] and it may also require considerable revision of the theory as its consequence. Nevertheless, when some indirect evidence of misspecification is available, researchers should not simply ignore it.

[26]Coltman et al. straightforwardly state that "One of the key operational issues in the use of formative indicators is that no simple, easy and universally accepted criteria exist for assessing their reliability" (Coltman et al. 2008, 1253).

References

Adamczyk, Amy, and Cassady Pitt. 2009 “Shaping Attitudes about Homosexuality: The role of Religion and Cultural context.” Social Science Research 38 (2): 338-351.

Alemán, José, and Dwayne Woods. 2016. "Value Orientations from the World Values Survey: How Comparable Are They Cross-Nationally?" *Comparative Political Studies* 49 (8): 1039-1067.

Alexander, Amy, Ronald Inglehart, and Christian Welzel. 2016. "Emancipating Sexuality: Breakthroughs into a Bulwark of Tradition." *Social Indicators Research* 129 (2): 909-935.

Asparouhov, Tihomir, Bengt Muthén, and Alexandre Morin. 2015. “Bayesian Structural Equation Modeling with Cross-Loadings and Residual Covariances: Comments on Stromeyer et al.” *Journal of Management* 41(6): 1561-1577.

Bowen. Donna Lee. 1997. “Abortion, Islam, and the 1994 Cairo Population Conference.” *International Journal of Middle East Studies* 29 (2): 161-184.

Brown, Timothy. 2006. *Confirmatory factor Analysis for Applied Research*. London, UK: The Guilford Press.

Browne, Michael, and Robert Cudeck. 1993. “Alternative Ways of Assessing Model Fit”. In *Testing Structural Equation Models*, eds. Kenneth A. Bollen and J. Scott Long. Newbury Park, CA: Sage, 136-162.

Byrne, Barbara, Richard Shavelson, and Bengt Muthén. 1989. “Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance.” *Psychological Bulletin* 105 (3): 456-466.

Chen, Fang Fang. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14 (3): 464-504.

Chen, Feinian, Patrick J. Curran, Kenneth A. Bollen, James Kirby, and Pamela Paxton. 2008. "An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models." *Sociological Methods & Research* 36 (4): 462-494.

Cheung, Gordon, and Roger Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9(2): 233-255.

Cieciuch, Jan, Eldad Davidov, Peter Schmidt, René Algesheimer, and Shalom H. Schwartz. 2014. "Comparing Results of an Exact versus an Approximate (Bayesian) Measurement Invariance Test: A Cross-Country Illustration with a New Scale to Measure 19 Human values." *Frontiers in Psychology* 5, 982 https://doi.org/10.3389/fpsyg.2014.00982.

Cieciuch, Jan, Eldad Davidov, Rene Algesheimer, and Peter Schmidt, P. 2017. "Testing for Approximate Measurement Invariance of Human Values in the European Social Survey." *Sociological Methods & Research*. Published online ahead of print April 10, 2017. https://doi.org/10.1177/0049124117701478

Coltman, Tim, Timothy Devinney, David Midgley, and Sunil Venaik. 2008. "Formative versus Reflective Measurement Models: Two Applications of Formative Measurement." *Journal of Business Research* 61 (12): 1250-1262.

Comrey, Andrew, and Howard Bee. 1992. *A First Course in Factor Analysis* (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Davidov, Eldad, Hermann Dülmer, Elmar Schlüter, Peter Schmidt, and Bart Meuleman. 2012. "Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance." *Journal of Cross-Cultural Psychology* 43 (4): 558-575.

Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40: 55-75.

Davidov, Eldad, Jan Cieciuch, Bart Meuleman, Peter Schmidt, René Algesheimer, and Mirjam Hausherr. 2015. "The Comparability of Measurements of Attitudes toward Immigration in the European Social Survey Exact versus Approximate Measurement Equivalence." *Public Opinion Quarterly* 79:S1 (January): 244-266.

Gelman, Andrew, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. 2013. *Bayesian Data Analysis*. Boca Raton: CRC Press.

Hoijtink, Herbert, and Rens van de Schoot. 2017. "Testing Small Variance Priors Using Prior-Posterior Predictive P-Values". *Psychological Methods*. Published online ahead of print April 3, 2017. http://dx.doi.org/10.1037/met0000131

Horn, John, and Jack McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18 (3): 117-144

Hu, Li-tze, and Peter Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling* 6 (1): 1-55.

Jackson, David, and Duane Alwin. 1980. "The Factor Analysis of Ipsative Measures." *Sociological Methods and Research* 9 (2): 218-238.

Inglehart, Ronald, Bi Puranen, and Christian Welzel. 2015. "Declining Willingness to Fight for One's Country: The Individual-Level Basis of the Long Peace."*Journal of Peace Research* 52 (4): 418-434.

Ippel, Lianne, John Gelissen, and Guy Moors. 2014. "Investigating Longitudinal and Cross Cultural Measurement Invariance of Inglehart's Short Post-Materialism Scale." *Social indicators research* 115 (3): 919-932.

Jackman, Simon. 2008. "Measurement." In: *Oxford Handbook of Political Methodology*, eds. Janet Box-Steffensmeier, Henry Brady, and David Collier. New York, NY: Oxford University Press, 119-51.

Jöreskog, Karl. 1971. "Simultaneous Factor Analysis in Several Populations." Psychometrika 36 (4): 409-426.

Meuleman, Bart. 2012. "When Are Item Intercept Differences Substantively Relevant in Measurement Invariance Testing? In: *Methods, Theories, and Empirical Applications in the Social Sciences*, eds. Samuel Salzborn, Eldad Davidov, and Jost Reinecke. Wiesbaden: VS Verlag für Sozialwissenschaften, 97-104.

Millsap, Roger E., and Jenn Yun-Tein. 2004. "Assessing Factorial Invariance in Ordered-Categorical Measures." *Multivariate Behavioral Research* 39 (3): 479-515.

Muthén, Bengt, and Tihomir Asparouhov. 2013. "Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory." *Psychological Methods* 17 (3): 313-335.

Muthén, Bengt, and Tihomir Asparouhov. 2013. "BSEM Measurement Invariance Analysis." *Mplus Web Notes* 17. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, Linda, and Bengt Muthén. 1998-2015. *Mplus User's Guide. 7th Edition*. Los Angeles, CA: Muthén and Muthén.

Oberski, Daniel. 2014. "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models." *Political Analysis* 22 (1): 45-60.

Schwartz, Shalom, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. "Refining the Theory of Basic Individual Values." *Journal of Personality and Social Psychology* 103 (4): 663-688.

Steenkamp, Jan-Benedict, and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25 (1): 78-107.

Stegmueller, Daniel. 2011. "Apples and Oranges? The Problem of Equivalence in Comparative Research." *Political Analysis* 19 (4): 471-487.

Tabachnick, Barbara, and Linda Fidell. 2007. *Using Multivariate Statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Van De Schoot, Rens, Anouck Kluytmans, Lars Tummers, Peter Lugtig, Joop Hox, and Bengt Muthén. 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

Van Vlimmeren, Eva, Guy Moors, and John Gelissen. 2017. "Clusters of Cultures: Diversity in Meaning of Family Value and Gender Role Items cross Europe." *Quality & Quantity* 51 (6): 2737-2760.

Vandenberg, Robert, and Charles Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4-70.

Welzel, Christian. 2013. *Freedom Rising*. New York, NY: Cambridge University Press.

Welzel, Christian, and Ronald Inglehart. 2016. "Misconceptions of Measurement Equivalence Time for a Paradigm Shift."*Comparative Political Studies*, 49 (8): 1068-1094.

Western, Bruce. 1999. "Bayesian Analysis for Sociologists: An Introduction." *Sociological Methods & Research* 28 (1): 7-34.

Wicherts, Jelte, and Conor Dolan. 2010. "Measurement Invariance in Confirmatory Factor Analysis: An Illustration Using IQ Test Performance of Minorities." *Educational Measurement: Issues and Practice* 29 (3): 39-47.

Zercher, Florian, Peter Schmidt, Jan Cieciuch, and Eldad Davidov. 2015. "The Comparability of the Universalism Value over Time and across Countries in the European Social Survey: Exact versus Approximate Measurement Invariance." *Frontiers in Psychology* 6, 733. https://doi.org/10.3389/fpsyg.2015.00733

Figures and Tables

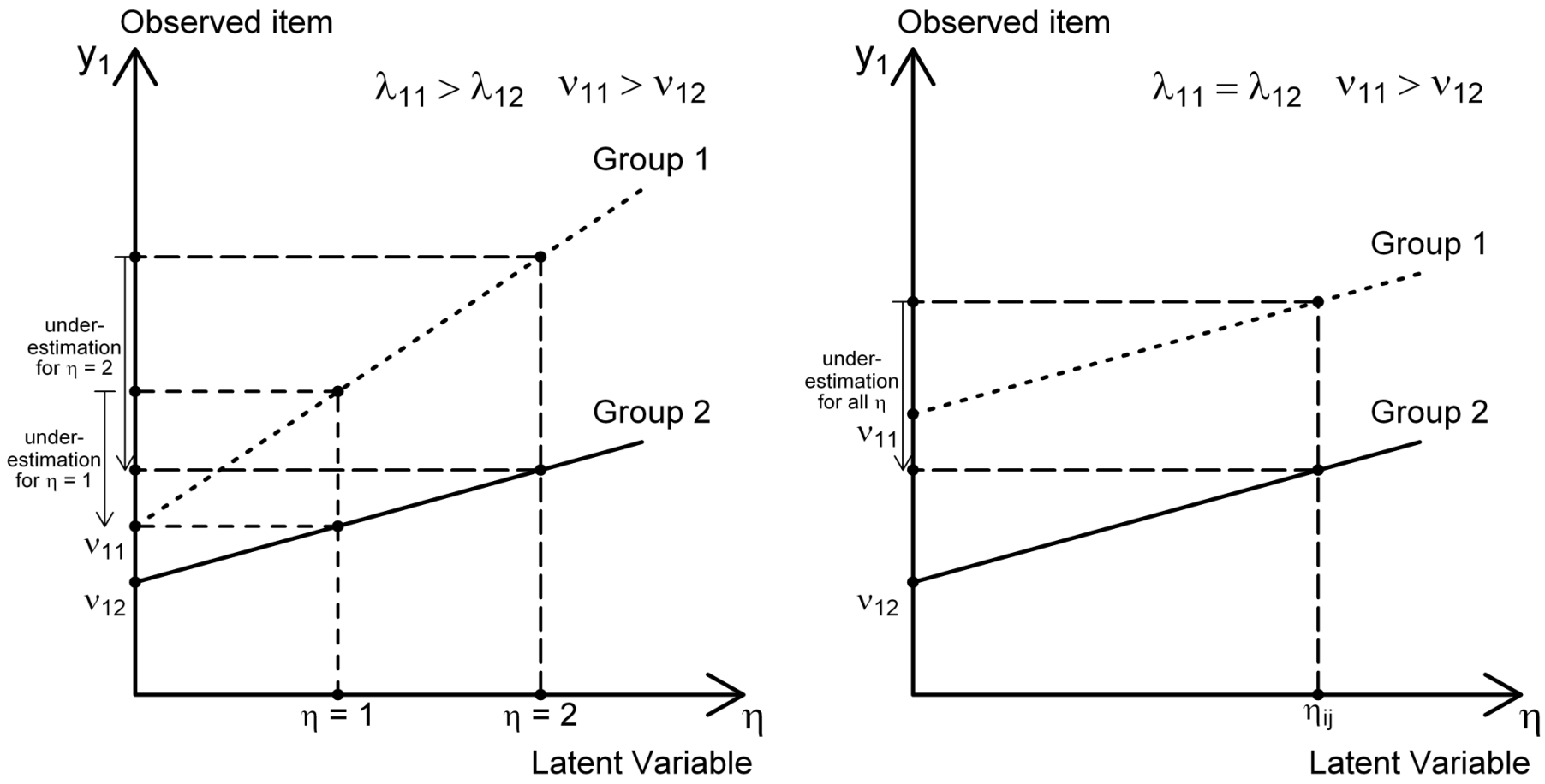

Figure B1. Regression lines for the prediction of values of the observed item y_1 on a latent variable η in two groups when intercepts and factor loadings are unequal (metric non-invariance; left panel) and when intercepts are unequal (scalar non-invariance; right panel).

Note: Similar figure appears in Wicherts and Dolan (2010).

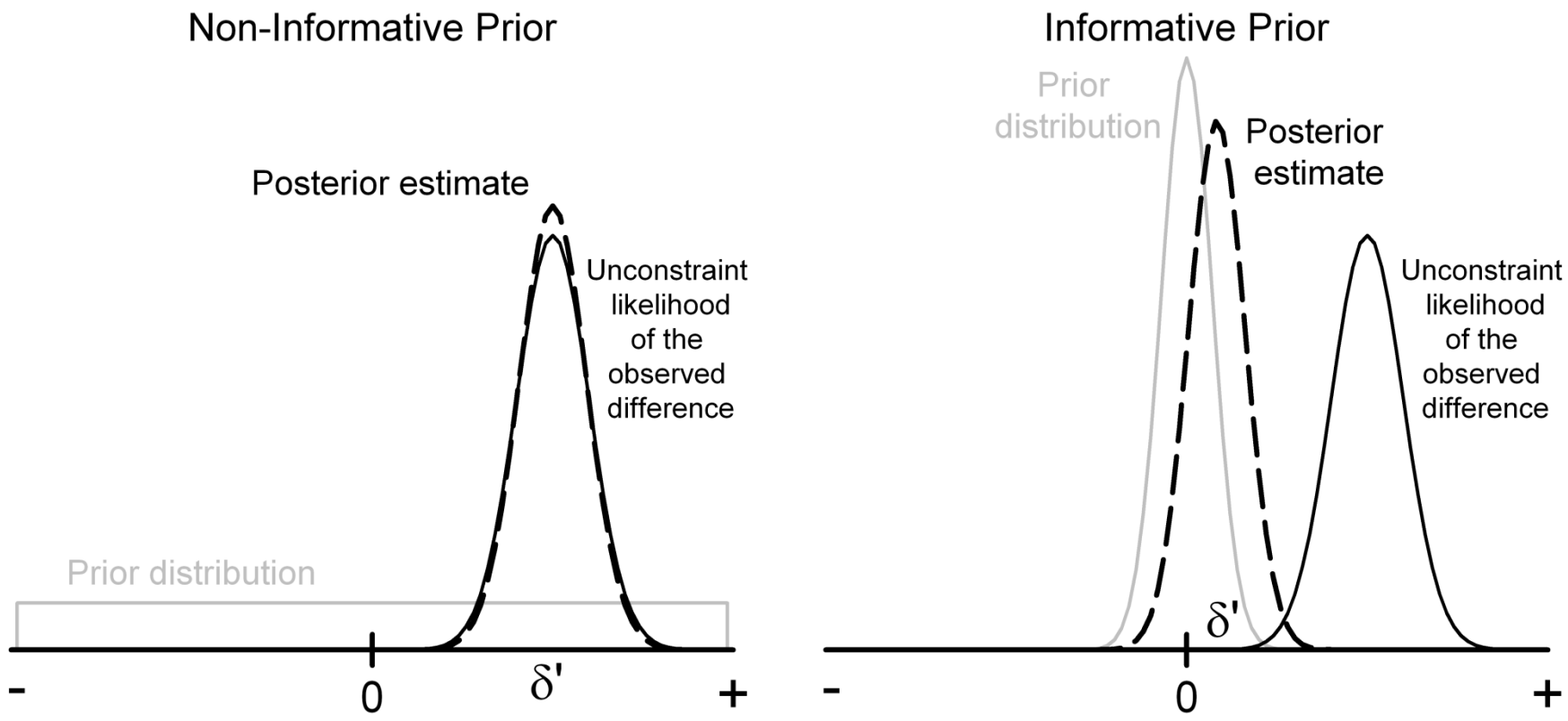

Figure B2. The influence of the prior on the posterior estimate of the difference in parameter values across groups.

*Note***:** Similar figure appears in van de Schoot et al. (2013).

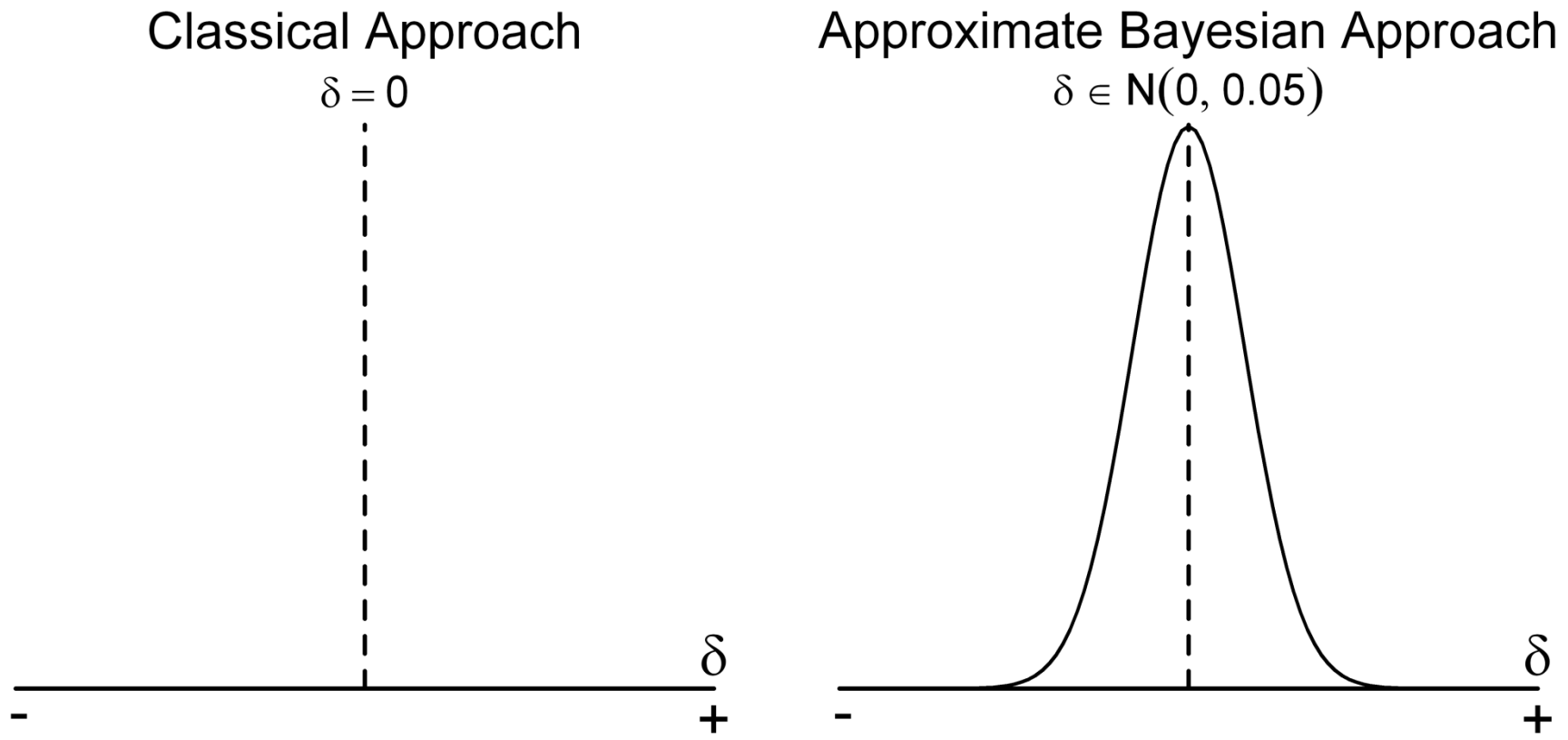

Figure B3. Prior treatment of the differences in parameters across groups in the classical (maximum likelihood) approach to invariance testing and the approximate Bayesian approach to invariance testing.

Note: Similar figures appear in Muthén and Asparouhov (2013) and Zercher et al. (2015).

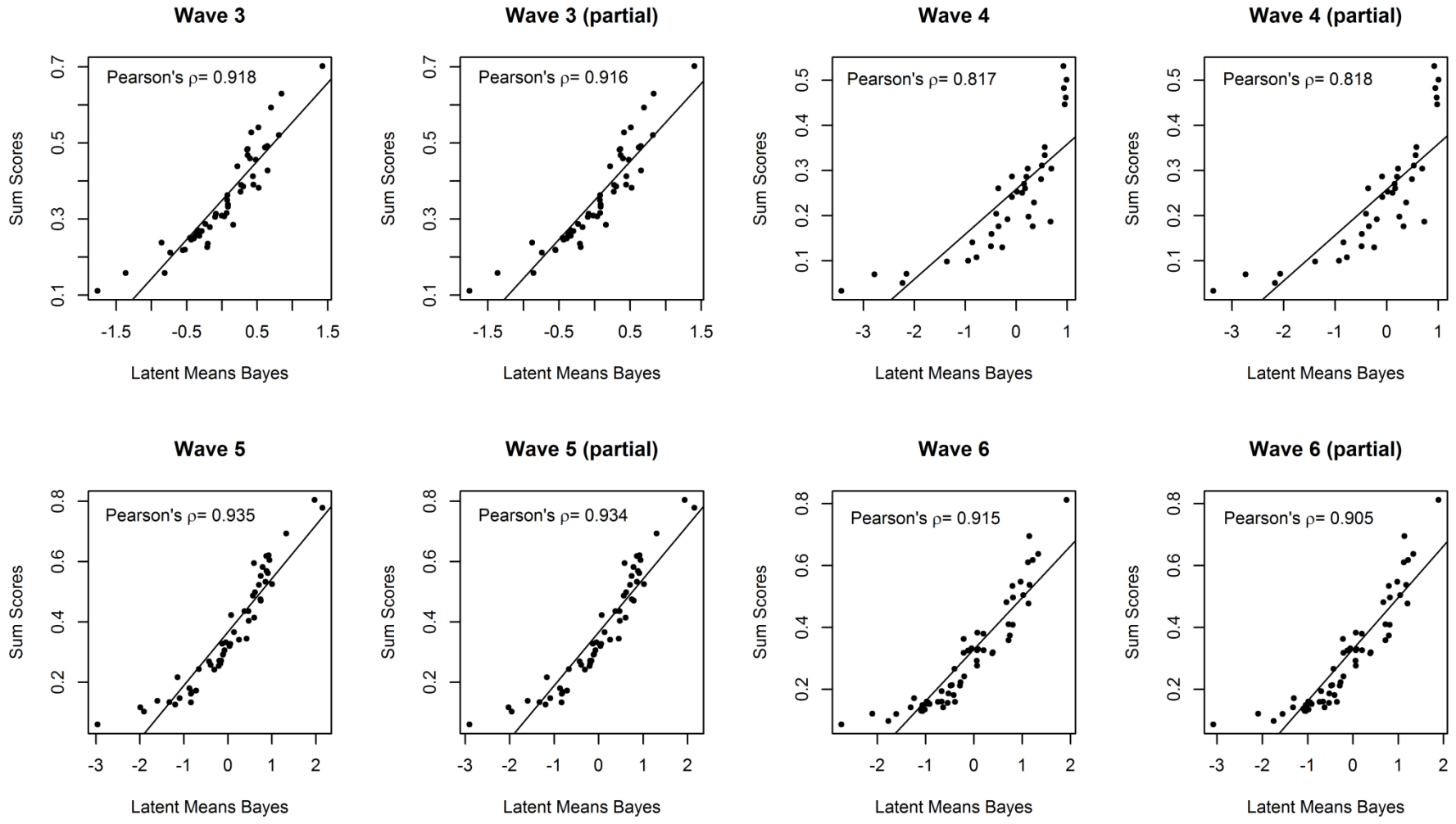

Figure B4. Relationship between the raw country mean scores for "choice" and the Bayesian country mean scores in WVS waves 3 to 6.

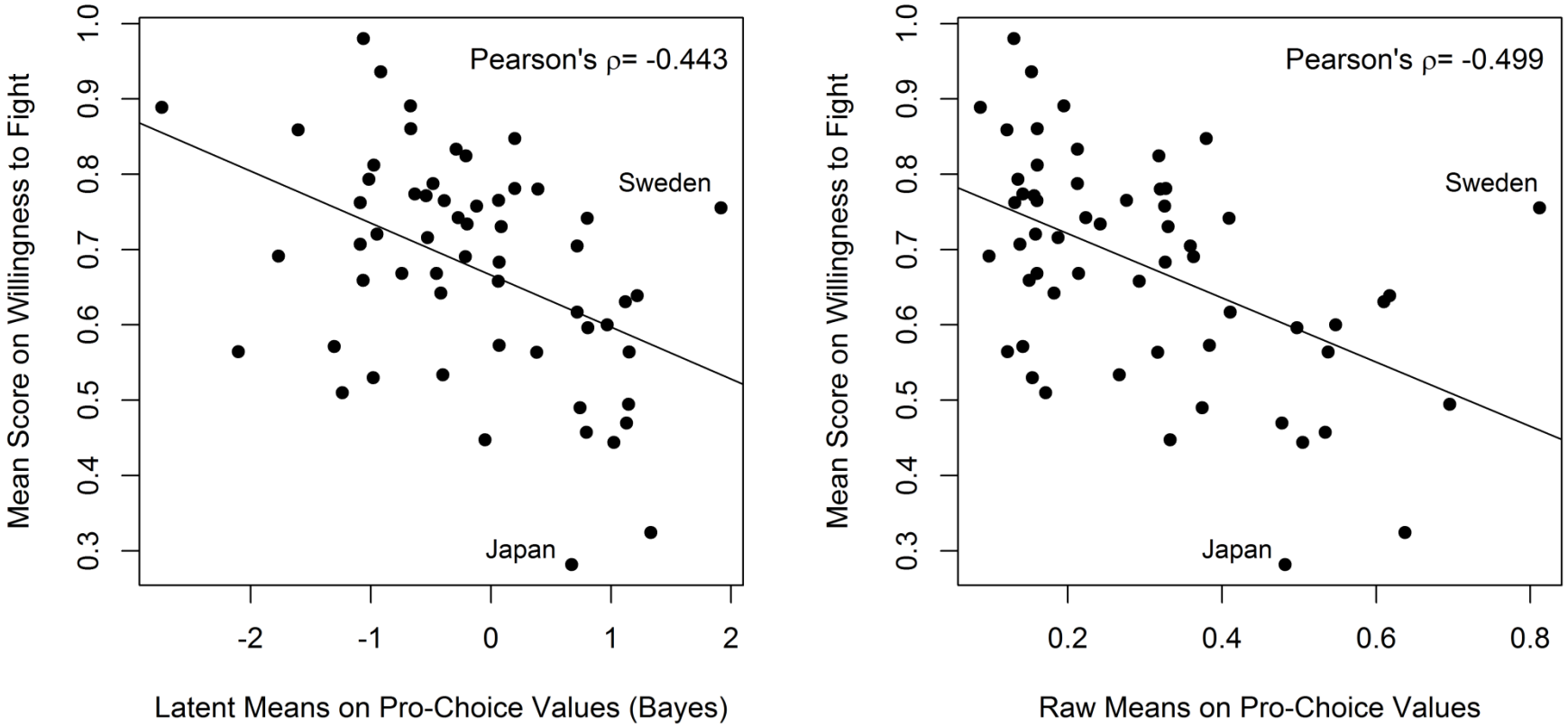

Figure D1. Association between the country mean scores on willingness to fight and (a) the country mean scores on "choice" based on the approximate Bayesian approach (Left Panel) and (b) the raw mean scores (Right Panel). Data for 58 countries from the 6 Wave of the WVS.

Figure E1: Panel A

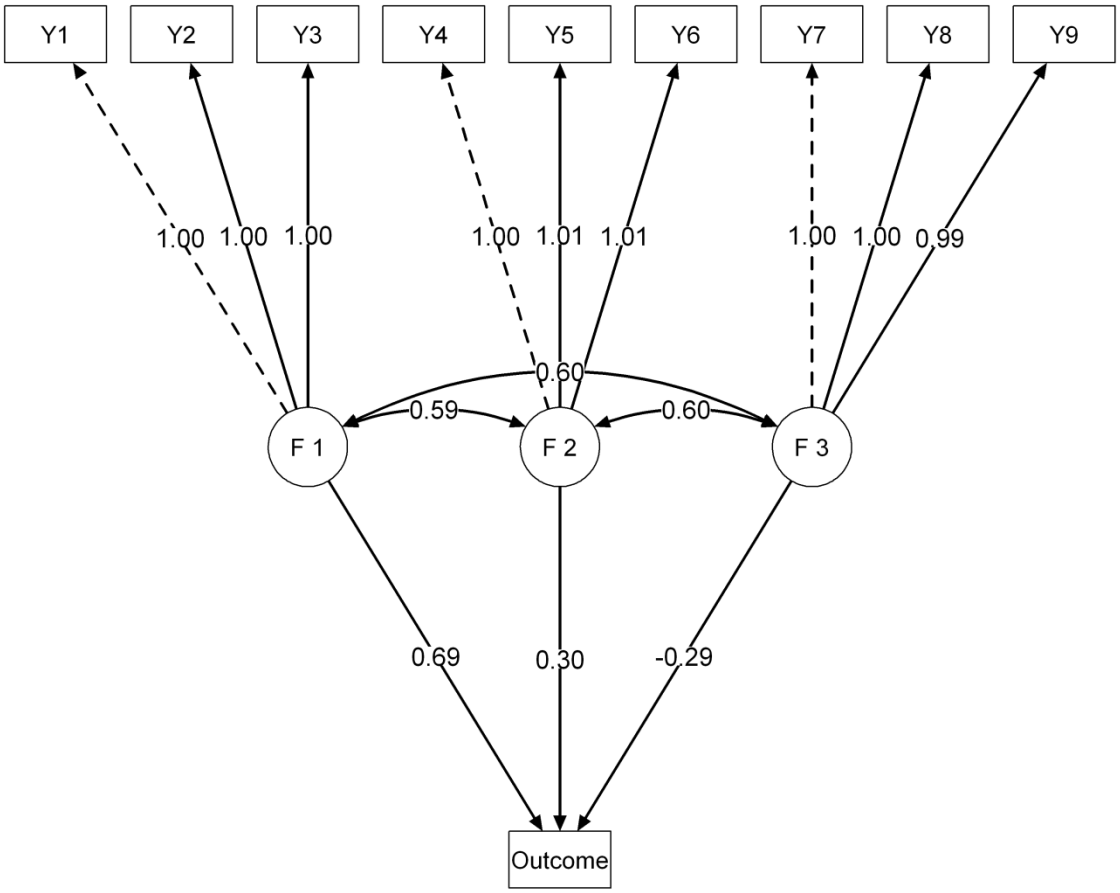

Figure E1. Illustration of incorrect inferences due to the use of a misspecified but well-fitting model

Notes: Panel A represents a true SEM-model with three latent variables (RMSEA = 0.000; CFI = 1, TLI = 1, the outcome's R^2 = 0.353); Panel B represents a misspecified SEM-model with a second-order latent variable (RMSEA = 0.045; CFI = 0.977, TLI = 0.968; R^2= 0.318); Panel C represents a regression model with an average score on all eight manifest variables as a predictor (RMSEA = 0.000; CFI = 1, TLI = 1, R^2= 0.198). Rectangles represent observed variables, ovals latent ones. An arc between latent variables represents their covariance. All parameter estimates shown are unstandardized values. All models are based on a simulated dataset (N = 100,000). The R package "simsem" (0.5-13) was used to simulate the data. Parameter values used for simulation are: all factor loadings = 1; variances of the outcome and latent variables = 1; all indicator residual variances = 1; covariance between latent factors = 0.6; regression coefficients are 0.7, 0.3 and −0.3 respectively.

Figure E1: Panel B

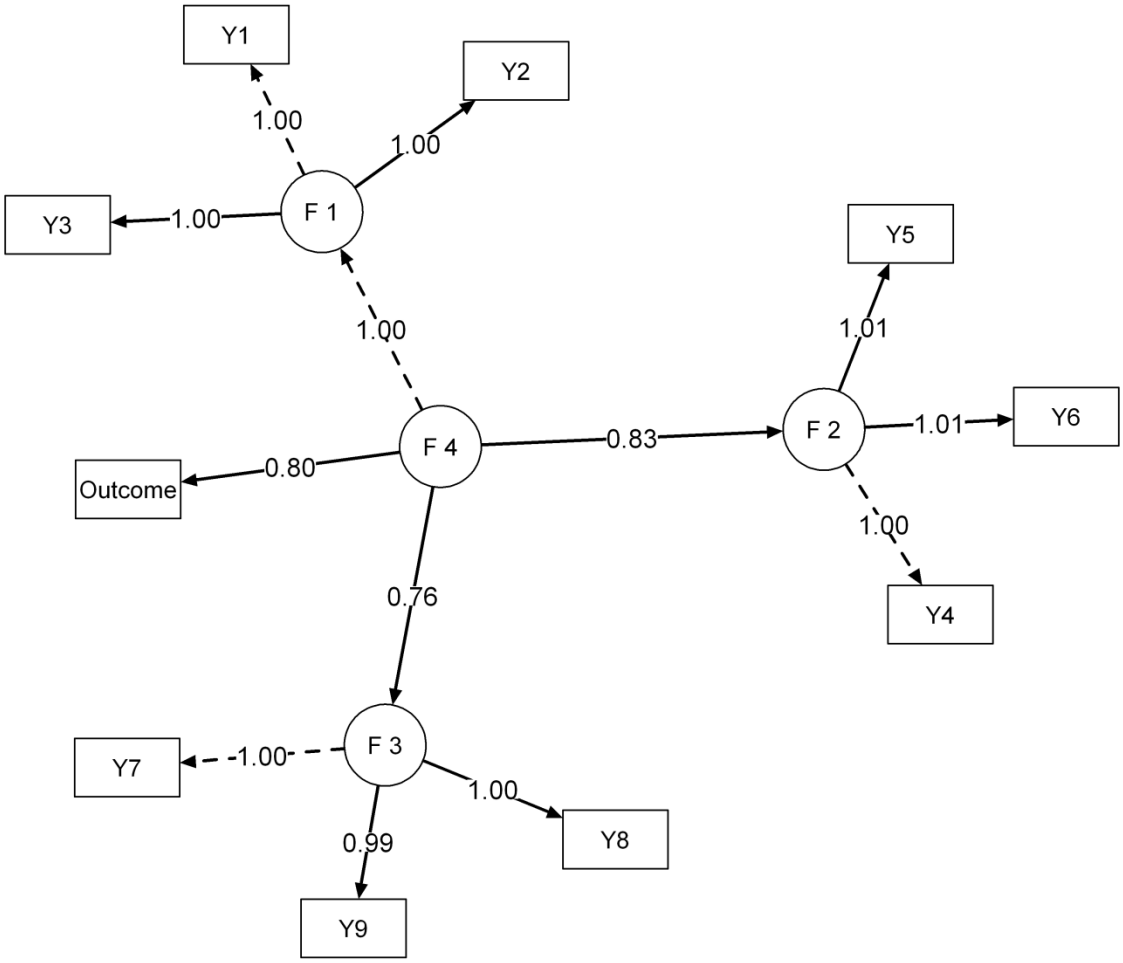

Figure E1: Panel C

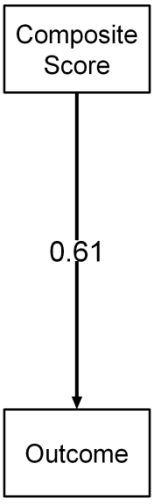

Table A1. CFA of 12 variables from WVS, 3rd-6th waves (1995–2014)

Variable	Factor	Pooled	Round 3	Round 4	Round 5	Round 6
Independence	Autonomy	0.489	0.574	0.561	0.453	0.445
Imagination	Autonomy	0.464	0.522	0.375	0.535	0.415
Obedience	Autonomy	0.545	0.563	0.586	0.557	0.479
Jobs	Equality	0.398	0.354	0.449	0.359	0.416
Leaders	Equality	0.821	0.725	0.812	0.853	0.865
Education	Equality	0.589	0.634	0.597	0.632	0.521
Homosexuality	Choice	0.833	0.787	0.741	0.865	0.856
Abortion	Choice	0.678	0.580	0.657	0.750	0.696
Divorce	Choice	0.705	0.668	0.664	0.737	0.711
Speech	Voice	0.341	0.355	**0.232**	0.406	**0.292**
Say_nat	Voice	**0.291**	0.429	0.367	**0.196**	**0.244**
Say_local	Voice	0.476	0.464	0.528	0.471	0.441
Autonomy	EVI	0.543	0.634	0.469	0.577	0.480
Equality	EVI	0.596	0.607	0.495	0.623	0.586
Choice	EVI	0.704	0.643	0.739	0.714	0.727
Voice	EVI	0.759	0.741	0.584	0.848	0.744
N		306406	77129	59030	83975	86272
CFI		0.905	0.856	0.897	0.925	0.897
TLI		0.874	0.810	0.864	0.901	0.864
RMSEA		0.041	0.049	0.042	0.040	0.042
P-value RMSEA < 0.05		1.000	0.849	1.000	1.000	1.000

Notes: Entries are standardized factor loadings. All estimates are significant at the 0.001 level. Loadings in bold are those lower than 0.30. Variable intercepts, thresholds and variances are not shown. Models were estimated in MPLUS version 7.11. National samples were weighted to equal size (N = 1,500). Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, the WLSMV estimator was used for parameter estimation. Pairwise present analysis was used to deal with missing values. N = number of observations used. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Standard Error of Approximation.

Table B1. Group-specific CFAs of 12 variables from the 6th wave of the WVS for ten cultural zones (2010–2014)

Variable	Factor	Islamic East	Indic East	Sinic East	Orthodox East	Old West	Reformed West	New West	Returned West	Latin America	sub-Saharan Africa
Independence	Autonomy	**0.268**	0.656	**0.109**	0.399	0.570	0.425	**0.284**	0.787	0.321	0.528
Imagination	Autonomy	0.324	**0.153**	0.455	**0.255**	0.429	0.489	0.391	0.413	**0.294**	0.342
Obedience	Autonomy	0.918	0.361	0.425	0.820	0.573	0.637	0.477	**0.079**	0.423	0.630
Jobs	Equality	0.546	0.340	0.336	0.366	**0.243**	–	–	–	**0.115**	0.337
Leaders	Equality	0.765	0.721	0.726	0.709	0.869	0.766	0.754	0.642	0.889	0.889
Education	Equality	**0.294**	0.589	0.722	0.569	0.648	0.848	0.783	0.704	0.581	0.430
Homosexuality	Choice	0.621	0.700	0.796	0.515	0.829	0.799	0.870	0.651	0.774	0.798
Abortion	Choice	0.875	0.778	0.660	0.766	0.720	0.781	0.721	0.793	0.558	0.908
Divorce	Choice	0.457	0.742	0.782	0.738	0.670	0.836	0.688	0.785	0.704	0.668
Speech	Voice	**0.247**	–	**0.299**	–	0.322	–	0.578	–	–	*-0.327*
Say_nat	Voice	**0.275**	–	0.443	0.407	**0.191**	0.464	*-0.391*	0.646	0.585	0.869
Say_local	Voice	0.486	1.00	0.582	0.435	0.568	0.699	–	**0.260**	0.401	**0.213**
Autonomy	EVI	0.416	n.s.	0.578	**n.s.**	0.303	0.712	0.832	0.554	0.444	0.525
Equality	EVI	0.511	**0.065**	0.523	**0.304**	0.394	0.614	0.560	0.419	**0.224**	**0.072**
Choice	EVI	0.336	n.s.	0.609	0.679	0.810	0.614	0.803	0.781	0.781	0.663
Voice	EVI	0.797	–	0.559	0.456	0.811	**0.266**	0.977	**n.s.**	0.328	**0.225**
N		18027	8453	8181	14842	2189	5154	4550	3568	11439	9869
CFI		0.848	0.902	0.893	0.917	0.910	0.953	0.961	0.875	0.966	0.913
TLI		0.800	0.867	0.859	0.886	0.881	0.932	0.943	0.818	0.953	0.885
RMSEA		0.044	0.040	0.042	0.036	0.045	0.038	0.033	0.052	0.024	0.034

Notes: Entries are standardized factor loadings. All estimates are significant at the 0.05 level (except those marked as n.s. = non-significant). Loadings in bold are those lower than 0.30. Negative loadings are in italic. Variable intercepts, thresholds and variances are not shown. Models were estimated in MPLUS version 7.11. National samples were weighted to equal size (N = 1,500). Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, the WLSMV estimator was used for parameter estimation. Pairwise present analysis was used to deal with missing values. N = number of observations used. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Standard Error of Approximation.

Table B2. Group-specific CFAs of 12 variables from the 5th wave of the WVS for ten cultural zones (2005-2009)

Variable	Factor	Islamic East	Indic East	Sinic East	Orthodox East	Old West	Reformed West	New West	Returned West	Latin America	sub-Saharan Africa
Independence	Autonomy	0.498	0.417	**0.202**	0.754	0.316	0.515	0.383	0.714	0.459	0.390
Imagination	Autonomy	0.452	**0.173**	0.382	**0.200**	0.622	0.597	0.504	0.308	0.498	0.353
Obedience	Autonomy	0.699	0.438	0.623	0.320	0.538	0.607	0.457	0.560	0.502	0.851
Jobs	Equality	0.675	**0.288**	**0.296**	**0.230**	**0.203**	**0.201**	–	**0.149**	**0.242**	0.401
Leaders	Equality	0.721	0.880	0.734	0.804	0.786	0.791	0.762	0.671	0.780	0.818
Education	Equality	0.440	0.614	0.656	0.538	0.784	0.762	0.725	0.772	0.688	0.597
Homosexuality	Choice	0.598	0.796	0.736	0.560	0.842	0.809	0.873	0.792	0.743	0.685
Abortion	Choice	0.814	0.925	0.712	0.844	0.771	0.739	0.700	0.800	0.584	0.938
Divorce	Choice	0.578	0.674	0.848	0.676	0.759	0.749	0.664	0.757	0.645	0.639
Speech	Voice	0.475	**0.279**	–	–	0.334	–	**0.176**	–	–	–
Say_nat	Voice	n.s.	–	0.506	0.490	0.308	0.500	0.351	0.638	–	–
Say_local	Voice	0.377	0.393	0.656	0.452	0.647	0.378	0.754	0.441	–	–
Autonomy	EVI	0.585	**0.257**	0.631	**0.286**	0.530	0.735	0.792	0.723	0.563	–
Equality	EVI	0.712	***-0.140***	0.428	**0.305**	0.554	0.603	0.635	0.531	**0.191**	–
Choice	EVI	0.404	0.433	0.601	0.575	0.735	0.734	0.718	0.726	0.725	–
Voice	EVI	0.551	0.819	0.429	0.402	0.670	**0.276**	0.352	**0.100**	–	–
N		12165	6751	8261	9576	5266	8438	5788	3044	12589	12097
CFI		0.936	0.921	0.920	0.926	0.911	0.966	0.860	0.949	0.981	0.967
TLI		0.915	0.891	0.890	0.898	0.882	0.953	0.807	0.930	0.971	0.950
RMSEA		0.030	0.042	0.039	0.031	0.051	0.032	0.058	0.036	0.022	0.030

Notes: Entries are standardized factor loadings. All estimates are significant at 0.05 level (except those marked as n.s. = non-significant). Loadings in bold are those lower than 0.30. Negative loadings are in italic. Variable intercepts, thresholds and variances are not shown. Models were estimated in MPLUS version 7.11. National samples were weighted to equal size (N = 1,500). Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, the WLSMV estimator was used for parameter estimation. Pairwise present analysis was used to deal with missing values N = number of observations used. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Standard Error of Approximation.

Table B3. Group-specific CFAs of 12 variables from the 4th wave of the WVS for ten cultural zones (1999-2004)

Variable	Factor	Islamic East	Indic East	Sinic East	Orthodox East	Old West	Reformed West	New West	Returned West	Latin America	sub-Saharan Africa
Independence	Autonomy	0.505	–	0.367	0.598	0.512	–	0.472	–	0.592	0.663
Imagination	Autonomy	0.418	–	0.309	**0.186**	0.552	–	0.483	–	0.425	**0.275**
Obedience	Autonomy	0.373	–	–	0.553	0.521	–	0.747	–	0.485	0.472
Jobs	Equality	0.470	0.318	0.460	**0.278**	**0.137**	–	**0.223**	–	**0.215**	0.412
Leaders	Equality	0.700	0.618	0.764	0.768	0.924	–	0.750	–	0.823	0.708
Education	Equality	0.515	0.661	0.643	0.731	0.699	–	0.752	–	0.644	0.665
Homosexuality	Choice	0.366	0.608	0.731	0.444	0.773	–	0.798	–	0.722	0.582
Abortion	Choice	0.786	0.822	0.543	0.767	0.801	–	0.720	–	0.561	0.802
Divorce	Choice	0.485	0.740	0.783	0.785	0.751	–	0.628	–	0.626	0.435
Speech	Voice	0.413	–	**0.075**	–	–	–	–	–	**0.270**	n.s.
Say_nat	Voice	–	0.424	0.476	0.639	**0.173**	–	–	–	**0.239**	–
Say_local	Voice	0.457	0.579	0.824	0.358	0.313	–	–	–	**0.281**	n.s.
Autonomy	EVI	–	–	0.550	0.484	0.582	–	0.588	–	0.493	0.472
Equality	EVI	–	–	0.556	0.487	0.455	–	0.463	–	0.305	**0.078**
Choice	EVI	–	–	0.699	0.451	0.756	–	0.872	–	0.633	0.545
Voice	EVI	–	–	0.605	**0.204**	0.356	–	–	–	0.869	n.s.
N		16516	9170	4562	7566	2408	–	3131	–	7436	8197
CFI		0.946	0.969	0.906	0.963	0.901	–	0.965	–	0.897	0.913
TLI		0.922	0.950	0.871	0.949	0.863	–	0.948	–	0.864	0.880
RMSEA		0.023	0.032	0.053	0.028	0.051	–	0.037	–	0.043	0.035

Notes: Entries are standardized factor loadings. All estimates are significant at 0.05 level (except those marked as n.s. = non-significant). Loadings in bold are those lower than 0.30. Negative loadings are in italic. Variable intercepts, thresholds and variances are not shown. Models were estimated in MPLUS version 7.11. National samples were weighted to equal size (N = 1,500). Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, the WLSMV estimator was used for parameter estimation. Pairwise present analysis was used to deal with missing values. N = number of observations used. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Standard Error of Approximation.

Table B4. Group-specific CFAs of 12 variables from the 3th wave of the WVS for ten cultural zones (1994-1998)

Variable	Factor	Islamic East (Turkey)	Indic East	Sinic East	Orthodox East	Old West (Spain)	Reformed West	New West	Returned West	Latin America	sub-Saharan Africa
Independence	Autonomy	0.797	0.712	0.403	0.715	0.762	0.527	0.476	0.481	0.671	0.523
Imagination	Autonomy	0.584	0.491	0.404	0.339	0.557	0.579	0.543	0.436	0.398	0.355
Obedience	Autonomy	0.452	0.654	0.708	0.468	0.521	0.608	0.508	0.719	0.489	0.593
Jobs	Equality	0.637	**0.266**	0.451	0.423	**0.362**	0.343	**0.277**	0.337	**0.245**	0.454
Leaders	Equality	0.499	0.871	0.593	0.599	0.691	0.718	0.744	0.549	0.713	0.598
Education	Equality	0.668	0.654	0.585	0.620	0.722	0.724	0.739	0.671	0.520	0.552
Homosexuality	Choice	NA	0.466	0.632	0.410	0.805	0.771	0.857	0.553	0.663	0.565
Abortion	Choice	NA	0.720	0.599	0.699	0.722	0.685	0.688	0.779	0.659	0.753
Divorce	Choice	NA	0.628	0.713	0.825	0.782	0.713	0.644	0.788	0.580	0.681
Speech	Voice	0.537	–	**0.268**	–	0.330	–	**0.061**	–	–	–
Say_nat	Voice	**0.150**	**0.242**	0.503	0.833	0.479	0.482	0.493	–	0.591	–
Say_local	Voice	0.388	0.936	0.751	0.339	0.447	0.443	0.615	–	0.400	–
Autonomy	EVI	0.808	0.882	0.669	0.408	**0.566**	0.830	0.807	0.594	0.693	0.492
Equality	EVI	0.753	0.345	0.449	0.465	0.572	0.683	0.680	0.551	**0.279**	0.387
Choice	EVI	NA	**0.154**	0.630	0.571	0.769	0.647	0.680	0.603	0.571	0.456
Voice	EVI	0.803	**0.250**	0.617	**0.198**	0.725	0.507	0.449	–	0.464	–
N		1907	5498	4583	19762	1211	7454	4791	9478	16714	4931
CFI		0.963	0.933	0.885	0.947	0.891	0.944	0.869	0.954	0.936	0.981
TLI		0.944	0.908	0.848	0.928	0.856	0.923	0.827	0.931	0.912	0.972
RMSEA		0.029	0.039	0.045	0.027	0.054	0.038	0.057	0.035	0.027	0.018

Notes: Entries are standardized factor loadings. All estimates are significant at 0.05 level (except those marked as n.s. = non-significant). Loadings in bold are those lower than 0.30. Negative loadings are in italic. Variable intercepts, thresholds and variances are not shown. Models were estimated in MPLUS version 7.11. National samples were weighted to equal size (N = 1,500). Due to the fact that 9 out of 12 observed indicators are categorical ordered variables, the WLSMV estimator was used for parameter estimation. Pairwise present analysis was used to deal with missing values. N = number of observations used. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Standard Error of Approximation. NA = Not Asked

Table B5. MGCFA for pro-choice values: deviations of loadings and intercepts from prior defined parameters (mean = 0, variance = 0.01) across ten cultural zones. WVS, 6th wave (2010-2014; 58 countries)

Parameter Type		Factor Loading		Intercept	
Item		Homosexuality	Abortion	Homosexuality	Abortion
Average	Parameter Value	0.891	1.143	0.019	0.000
Standard Deviation of the Average Parameter Value		0.007	0.008	0.005	0.005
Cultural Zone	Islamic East		0.352	-0.310	
	Indic East			0.233	0.243
	Sinic East		-0.218		
	Orthodox East	-0.457		-0.411	
	Old West				-0.279
	Reformed West			0.388	
	New West	0.302			
	Returned West				
	Latin America		-0.455		-0.369
	sub-Saharan Africa				0.306

Notes: Entries are unstandardized parameter estimates for the rows 1-2 and unstandardized deviations of zone-specific factor loadings and intercepts from the population average values for the rows 3-12. All estimates are significant at the 0.05 level. Only deviations out of $[-0.2; 0.2]$ range are shown. All manifest variables were centered before the analysis; that is why the estimated intercepts for both items are very close to zero. Item "divorce" were used as the marker variable; the factor loading for that item were constrained to 1, and the intercept were constrained to 0 in all cultural zones.

Table B6. Model fit coefficients of Bayesian MGCFA for pro-choice values for each WVS wave (prior variance = 0.05).

Wave	Approximate Invariance Model		Partial Approximate Invariance Model		Number of Countries
	PPP	χ^2Credibility Interval	PPP	χ^2Credibility Interval	
Wave 1	0.466	[-32.348; 36.406]	–	–	8 countries*
Wave 2	0.326	[-39.100; 62.504]	–	–	18 countries
Wave 3	0.265	[-56.688; 114.148]	0.341	[-67.188; 103.632]	51 countries
Wave 4	0.013	[10.992; 157.096]	0.038	[-5.747; 137.409]	37 countries*
Wave 4 reduced	0.214	[-37,268; 96.003]	0.355	[-51.122; 80.160]	33 countries: Saudi, Bangladesh, Pakistan, and Algeria excluded
Wave 5	0.106	[-32.144; 140.523]	0.146	[-39.440; 134.419]	54 countries*
Wave 6	0.034	[-5.370; 179.934]	0.081	[-27.182; 154.933]	58 countries*
Wave 6 reduced	0.171	[-46.358; 130.316]	0.222	[-53.728; 121.470]	52 Countries: Bahrain, Palestine, Jordan, Lebanon, Morocco, and Pakistan excluded

Notes: PPP = posterior predictive p-value; χ^2Credibility Interval = 95% credibility interval for the difference between the observed and the replicated chi-square values. Item measuring people's acceptance of divorce is used as the marker variable. The mean of the differences in loadings and intercepts across countries is defined as zero and the variance of these differences as 0.05. Models were estimated in MPLUS version 7.11. With Bayesian analysis, modeling with missing data gives asymptotically the same results as full information maximum likelihood estimation under missing at random (MAR) mechanism. (Muthén and Muthén, 1998-2015: 386).Traditional diagnostic tools for Bayesian analysis, such as trace and autocorrelation parameter plots, Gelman-Rubin potential scale reduction factor and Kolmogorov-Smirnoff test, indicate good convergence for all the models presented.

*Some countries are not included in the MGCFA for that wave because one or more items measuring pro-choice orientations were not asked in those countries.

Table B7. Country-specific factor loadings and intercepts for the most dissimilar countries in the WVS waves 4 and 6.

Country	Cultural Zone	PPP	χ^2Credibility Interval	Homosexuality Loading	Abortion Loading	Homosexuality intercept	Abortion Intercept
Wave 4		**0.013**	**[10.992; 157.096]**	**0.632**	**1.070**	**-0.091**	**-0.002**
Algeria	Islamic East	0.063	[-3.230; 25.826]	0.704	1.280	-0.631	-0.706
Bangladesh	Indic East	0.002	[10.117; 38.488]	0.222	1.208	-0.340	0.474
Pakistan	Indic East	0.178	[-6.257; 17.349]	0.013	0.864	-0.520	0.319
Saudi Arabia	Islamic East	0.057	[-2.876; 25.683]	0.238	1.459	-0.432	-0.305
Wave 6		**0.034**	**[-5.730;179.934]**	**0.753**	**1.135**	**-0.083**	**0.008**
Bahrain	Islamic East	0.153	[-6.753;22.339]	1.273	1.475	0.181	0.269
Palestine	Islamic East	0.135	[-6.014;20.227]	0.419	1.420	-0.416	0.238
Jordan	Islamic East	0.186	[-7.181;18.760]	0.378	1.337	-0.478	0.184
Lebanon	Islamic East	0.134	[-5.991;21.685]	1.105	1.475	0.087	0.201
Morocco	Islamic East	0.087	[-4.094;23.691]	0.487	1.516	-0.612	-0.340
Pakistan	Indic East	0.039	[-1.388;28.645]	1.125	1.254	0.294	0.326

Notes: The factor loading for the item measuring people's acceptance of divorce is fixed to 1 and the intercept for the same item is fixed to 0 in all groups in order to identify the model. Only countries with group-specific PPP lower than 0.2 are shown. In bold are fit indices and the average values of the respective parameters for the overall model for each wave. Note that MPLUS 7.11 does not provide standardized sample-average parameter estimates so unstandardized group-specific parameter values are used for comparisons.

Table B8. Number of countries and respondents included in the analysis by WVS wave.

Wave	Number of Countries Covered	Number of Countries Included In the Analysis	Countries Excluded	Number of Respondents Included In the Analysis
Wave 1	10	8	Sweden, USA	9,924
Wave 2	18	18		24,308
Wave 3	54	51	Bangladesh, Pakistan, Turkey	72,129
Wave 4	41	37	Iraq, Morocco, Turkey	51,534
Wave 5	58	54	Iraq, Morocco, Peru, Egypt	73,878
Wave 6	60	58	Kuwait, Egypt	81,972

Notes: Countries listed in the column "Countries excluded" were not included in the Bayesian MGCFA for the respective wave because one or more items measuring pro-choice orientations were not included in the national WVS questionnaires in those countries. More descriptive statistics and the full lists of countries covered in each wave are available at www.worldvaluessurveys.org

Table B9. Model fit coefficients of Bayesian MGCFA for pro-choice values with stricter prior levels of invariance.

Wave	Prior Variance	PPP	χ^2 Credibility Interval	Full/Partial Invariance	Number of Countries
Wave 1	0.01	0.035	[-2.644; 72.046]	Full	8 countries*
Wave 2	0.02	0.040	[-5.631; 107.171]	Full	18 countries
Wave 3	0.02	0.057	[-12.971; 158.907]	Partial	51 countries*
Wave 4	0.02	0.068	[-15.751; 128.806]	Partial	33 countries*: Saudi, Bangladesh, Pakistan and Algeria excluded
Wave 5	0.03	0.029	[-1.841; 153.035]	Partial	54 countries*
Wave 6	0.03	0.065	[-18.518; 159.864]	Partial	52 Countries*: Bahrain, Palestine, Jordan, Lebanon, Morocco, and Pakistan excluded

Notes: PPP = posterior predictive p-value; χ^2 Credibility Interval = 95% credibility interval for the difference between the observed and the replicated chi-square values. Item measuring people's acceptance of divorce is used as the marker variable. Model fit coefficients are shown for the models with the lowest prior variance that does not result in the zero PPP for the respective wave. With Bayesian analysis, modeling with missing data gives asymptotically the same results as full information maximum likelihood estimation under missing at random (MAR) mechanism. Traditional diagnostic tools for Bayesian analysis, such as trace and autocorrelation parameter plots, Gelman-Rubin potential scale reduction factor and Kolmogorov-Smirnoff test, indicate good convergence for all the models presented.

*Some countries are not included in the MGCFA for that wave because one or more items measuring pro-choice orientations were not asked in those countries.