Comparability of Formative Measures in Cross-National Surveys

Boris Sokolov
bssokolov@gmail.com

LCSR HSE

25.05.2021

Intro

- Motivation:
  - Attitudes, values, beliefs etc. are difficult to define and even more difficult to measure
  - Especially in comparative contexts
  - Relevant issues and pitfalls are rarely addressed (and even recognized) by many applied researchers
  - Long-standing debate on what cross-national comparability is and how to achieve/prove it.
- Contribution:
  - Special focus on the use of so called formative measures in cross-cultural studies
  - Example: Liberal and authoritarian notions of democracy (LNDs/ANDs; see Kirsch and Welzel 2019) from the World Values Survey
  - Thanks to Cristian Welzel and late Ronald Inglehart for the inspiration and for the ideas.

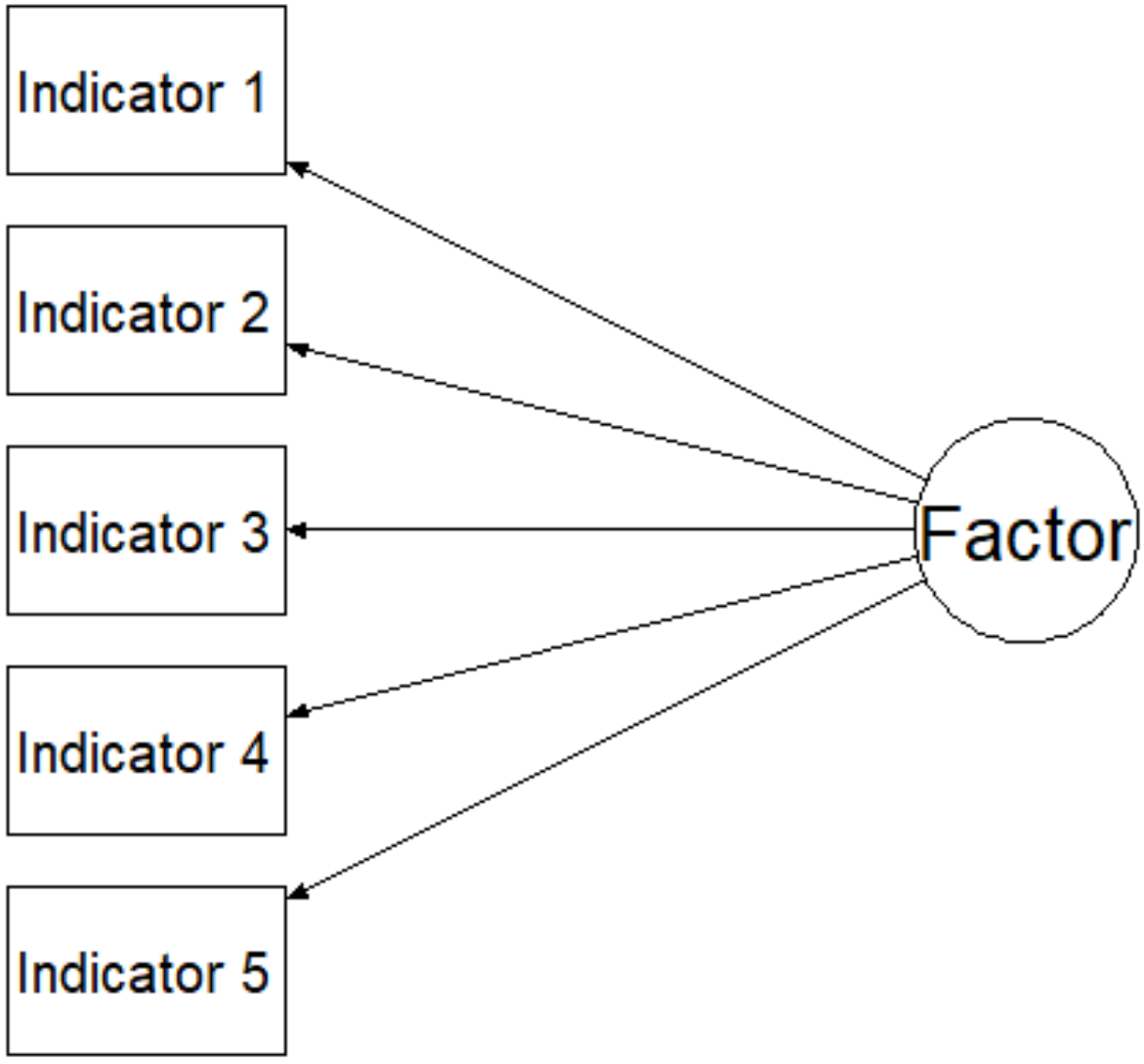
What LNDs and ANDs are?

- Question: "Many things are desirable, but not all of them are essential characteristics of democracy. Please tell me for each of the following things how essential you think it is as a characteristic of democracy. Use this scale where **1** means *"not at all an essential characteristic of democracy"* and **10** means it definitely is *"an essential characteristic of democracy""*
- Liberal notions of democracy (LNDs):
  - People choose their leaders in free elections (**V133** in the WVS-6 questionnaire)
  - Civil rights protect people from state oppression (**V136**)
  - Women have the same rights as men (**V139**)
- Authoritarian notions of democracy (ANDs):
  - Religious authorities ultimately interpret the laws (**V132**)
  - The army takes over when government is incompetent (**V135**)
  - People obey their rulers (**V138**)

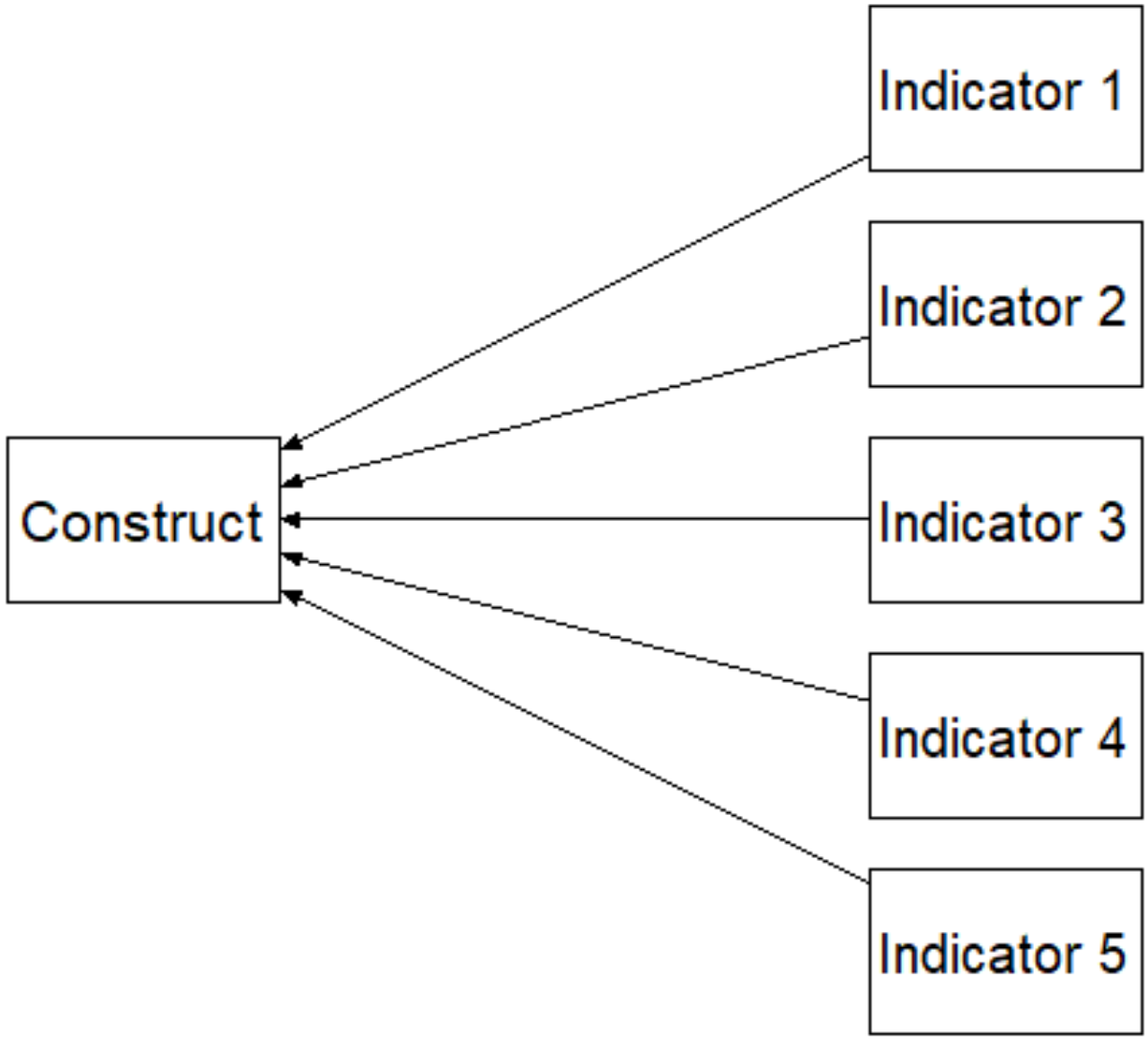
Why LNDs and ANDs?

- Related to important recent results in comparative politics
  - Why people widely support democracy in countries where it is actually absent? Which implications this paradoxical support has for the persistence of authoritorian rule? (Kircsh and Welzel 2019; Kruse, Ravlik, and Welzel 2019; Zagrebina 2019; Claassen 2020)
- Were not tested for comparability using tools that are considered standard by most methodologists today (but see Ariely and Davidov 2011)
- Were justified by the authors using an alternative interpretation of cross-national comparability.
- Nice opportunity to illustrate, with a practical example, modern approaches to comparability assessment and, more generally, construct validation and related conceptual and methodological issues.

Reflective constructs



Formative constructs



Formative vs. Reflective measures

- Direction of causation: from indicators to construct (F) vs. from construct to indicators (R)
  - Construct exists at the same level (F) or at a deeper level (R) of abstraction than its indicators
  - construct is formed by its indicators (F) or it exists independently of its indicators (R)
- Strength of correlations between indicators of a construct: any (F) or high (R)
- Degree of interchangeability among indicators: any (F) or high (R)
- Nomological nets of indicators: can be different (F) or same (strongly overlapping) (R)
- Statistical interpretation: constructs are linear functions/composites of indicators (F) or latent factors responsible for covariation between indicators (R)
- Model quality: Predictive power (F) or inter-item correlations (R)as a source of model validity

Formative vs. Reflective measures

- Relective model (MGCFA: $i-$ individuals, $j-$ items, $g-$ groups):

$$y_{ijg} = \nu_{jg} + \lambda_{jg} \times \eta_{ig} + \varepsilon_{ijg}$$

- Formative model 1:

$$\eta_{ig} = \beta_{1g} \times x_{1ig} + \beta_{2g} \times x_{2ig} + \beta_{3g} \times x_{3ig} + \delta_{ig}$$

- Formative model 2:

$$\eta_{ig} = w_{1g} \times x_{1ig} + w_{2g} \times x_{2ig} + w_{3g} \times x_{31ig}$$

- Formative model 3 (e.g., ANDs):

$$\eta_{ig} = 1 \times x_{1ig} + 1 \times x_{2ig} + 1 \times x_{3ig}$$

Comparability in the reflective approach

- Comparability as measurement invariance (MI):
  - Same model measures same concept under different circumstances
  - Configural MI: Similar factor structures in different countries $\rightarrow$ similar construct contents $\rightarrow$ numeric comparisons are premature
  - Metric MI: Equivalent factor loadings across countries $\rightarrow$ equivalent latent measurement units $\rightarrow$ latent variances and covariances are comparable
  - Scalar MI: Equivalent item intercepts across countries $\rightarrow$ equivalent latent scale orignis $\rightarrow$ laten means are comparable
- Problems:
  - Equivalence of numeric scores does not prove conceptual or functional equivalence (and vice versa)
  - Rarely holds in practice (but see *partial MI* and *approximate MI*)
  - Some constructs do not require strong inter-item correlations by definition

Summary of reflective analyses of LNDs and ANDs

- WVS-6: 60 countries, $\approx$ 90K respondents
- Exploratory analysis (not shown)
- Confirmatory factor analysis of LNDs and ANDs (see Appendix III)
- Basic and advanced MGCFA invariance tests for LNDs (see Appendix III)

Summary of reflective analyses of LNDs and ANDs

- LNDs: configural and partial metric, but not scalar MI
- ANDs: not even configurally invariant
- Latent LND means can be recovered quite precisely using the novel MGCFA alignment approach
- Seems that LNDs are reasonably comparable across most WVS-6 countries.
- What about ANDs? Should we abandon that measure?

Problems with formative models

- Models 1 and 2:
  - Difficult to identify (typically require additional reflective indicators) and estimate statistically;
  - Interpretational confounding (parameter values may depend on the choice of identifying reflective variables)
  - Many other (Wilcox et al. 2008; Bollen and Diamantopoulos 2017)
- Model 3: makes very strong assumptions
  - No measurement error at either indicator or construct level
  - Equal contributions of every indicator to the total score
  - Equal measurement scales and reference points across nations
  - How to assess comparability?

Normative/theoretical benchmarking argument

- Value and attitudinal constructs encoded in modern cross-national surveys are typically of the Western origin
- Some scholars and activists think that it is (ethically) bad.
- It may also undermine comparablity: non-Western respondents may either misunderstand to some extent, or non understand at all, what they are asked about
  - respective constructs may well not exist as collective ideational entities outside the "global" West.
- Yet, if one believes that *ideas* may influence actual social and political processes, it is then natural to assert that what matters is not only the degree of commitment to some (equivalently perceived) idea but also the degree of (mis-)understanding of the idea: hard to truly commit to what you don't even understand.
- The Western (*Western academic*) roots of the concept od democracy provide a promising benchmark for ensuring comparability:
  - Democracy is what (most) political scientists have agreed upon it is: **Clear and invariant measurement reference point**
  - If you disagree with a given operational definition of democracy you are either (a) a political scientist/philosopher or (b) simply misunderstand the term

Compositional substitutability

- Formative measures do not require strong inter-item correlations $\Longrightarrow$ no way to test model quality via conventional SEM approaches
- Comparability tests for formative constructs are somewhat conceptually self-contradicting (mimic reflective concepts of configural, metric and scalar MI: Diamantopoulos and Papadopoulos 2010; Henseler et al. 2016)
- Compositional substitutability criteria of measurement validity: Welzel and Inglehart CPS 2016 , Welzel et al SMR 2021
- Nomological validity: strong correlations between a formative measure and its expected correlates
- In terms of predictive power, the overall score overperforms its specific indicators
  - Naturally incorporates cross-national differences in a relative salience of indicators $\Longrightarrow$ allows to capture the impact of different national historical legacies on political beliefs
- Cross-national difference in inter-item alignment do not affect the predictive performance of the overall construct
- Personally, I do not think that the latter condition is relevant $\Longrightarrow$ other ways to deal with measurement error may be more promising
- CS approach works better at the aggregate level rather than at the individual one, but national cultures are meaningful analytic units

ANDs: Formative or Reflective?

- Direction of causality:
  - Cross-level: from culture to personal attitudes
  - Individual-level: from personal latent understanding of democracy to responses (R)
  - Country-level: from particular dimensions to the construct (F)
- Interchangeability
  - Reasonable for LNDs but not ANDs
  - Confusing democracy with military rule may have nothing to do with confusing democracy with theocracy.
  - Sharing both misunderstandings probably reflect a higher level of misunderstanding that endorsing only one: Additive effect, not interchangeability

ANDs: Formative or Reflective?

- Covariation:
  - Understanding democracy in a correct way requires that one could correctly identify all key components of the concept, so some, and even large covariance between the LND indicators is to be expected
  - But there multiple possible ways to misunderstand democracy, and I see no reasons why they should covary (though they can in some countries, of course, as it can be seen from the data)
  - Paraphrasing Leo Tolstoy, it can be claimed that all democratic people (and countries) have the same understanding of what democracy is, but all non-democratic peoples (and countries) have their own concept of democracy.
- Nomological net:
  - Some antecedents are similar but not all (see evidence below)
- Formative interpretation of ANDs seem plausible, BUT...

Formative measurement errors

- The way ANDs are measured (mean or sum score over three items) resembles *Formative model 3*
- *Model 3* assumes perfect correspondence between recorded responses and unobserved true item scores and equal item weights.
- It also assumes that there is neither random nor systematic measurement error in indicators and the resulting construct
- This assumption is unlikely to hold:
  - Some other misunderstandings are possible but not reflected in WVS $\implies$ construct-level ME
  - Random or systematic noise in individual responses

Aggregation may help with random ME

- $x^*_{ij} = x_{ij} + \varepsilon_{ij} + u_j$
  (a) i indexes individuals and j indexes countries
  (b) x^*_{ij} is the observed score and x_{ij} is the true score
  (c) $\varepsilon_i \sim \mathcal{N}(0, \sigma_{\varepsilon_i})$ is random ME
  (d) u_j is systematic ME, constant across individual within countries
- $\overline{x}^*_j = \mathbb{E}[x_{ij} + \varepsilon_{ij} + u_j] = \overline{x}_j + u_j$
- Aggregation removes random individual ME but not systematic (country-specific) ME
- In cross-national data, u can be seen as a country-level error, which can be random (uncorrelated with any other relevant variable) or systematic (when u correlates with substantive CL variables).
- How critical is the country-level error u for substantive scientific goals?

Country-level error and mean comparisons

- Suppose that $Cov(\overline{x}_j, u_j) = 0$ (no correlation between true means on x and u)
- If we are interested in comparison of means u increases the variance in $\overline{x}_j^*$ compared to $\overline{x}_j$
  - $Var(\overline{x}_j^*) = Var(\overline{x}_j + u_j) = Var(\overline{x}_j) + Var(u_j)$
- u may also distort the ranking of countries on $\overline{x}_j^*$, compared to that on $\overline{x}_j$
- Rankings of both reflective and formative means are not so accurate in the presense of ME (easy to see when comparing LND reflective means obtained using different MGCFA approaches; see Appendix).
- What about regression estimates?

Country-level error and regression estimates

- True model: $y_j = \beta_0 + \beta_1 \times \overline{x}_j + \varepsilon_j$
- Estimated model: $y_j = \beta_0 + \beta_1^* \times \overline{x}_j^* + \varepsilon_j$
- Let
  (a) $Cov(\overline{x}_j, u_j) = 0$,
  (b) $\varepsilon_j \sim \mathcal{N}(0, \sigma_{\varepsilon_j})$,
  (c) $Cov(\varepsilon_j, \overline{x}_j) = 0$,
  (d) $Cov(\varepsilon_j, u_j) = 0$
  (e) $Cov(y_j, u_j) = 0$.
- $$\beta_1^* = \lambda \beta_1$$

  where $0 \leq \lambda \leq 1$ ($\lambda -$ reliability/attenuation ratio)
- The most likely bias in β is *toward* the null, not *away*
- Significant βs are likely to be underestimated under ME in x, not overestimated: so no reasons to criticize respective theories on the grounds of ME-driven false positives
- Other problems: (1) biased intercept; (2) inflated residual variance; (3) reduced power (Greenwood 2012)

Complications

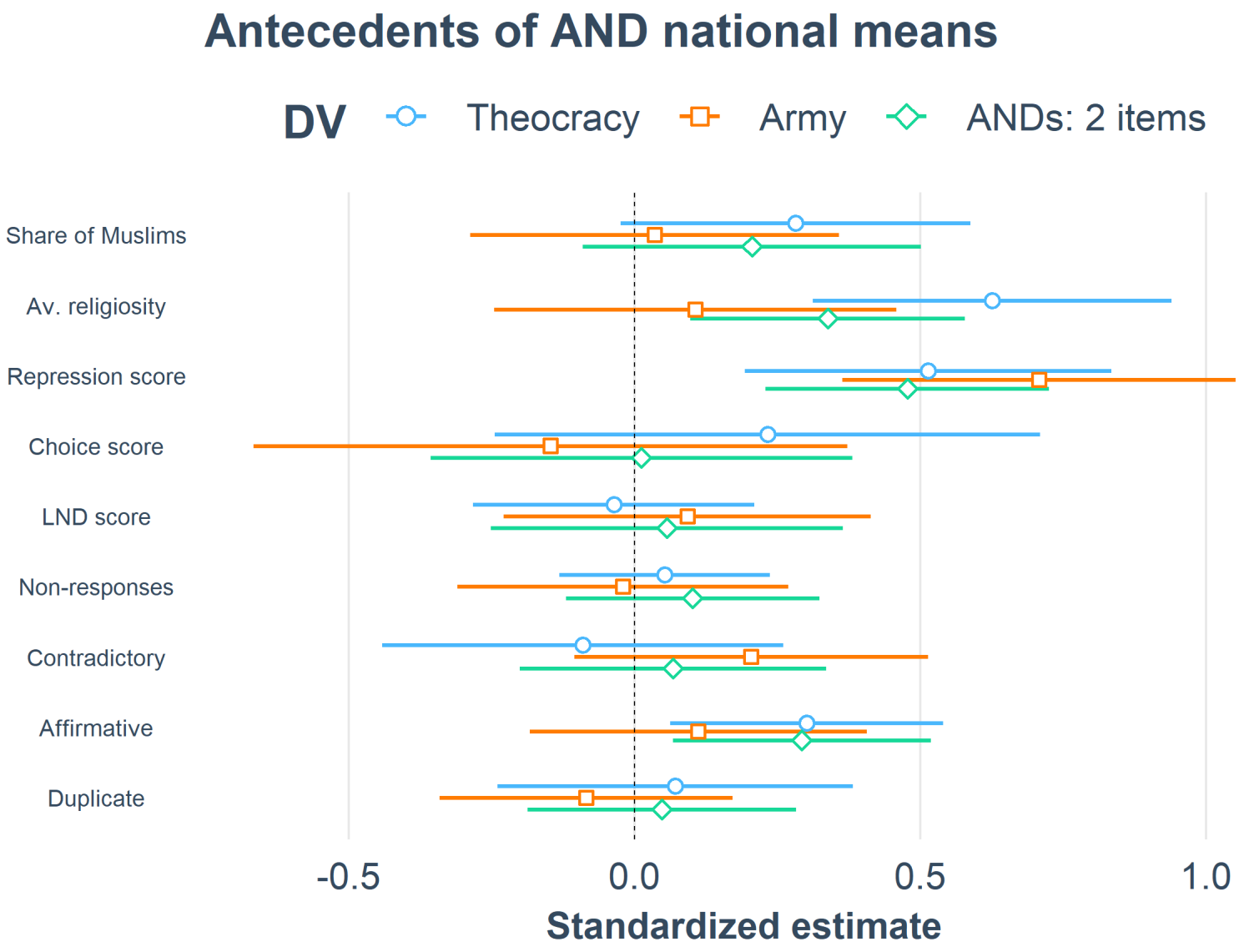
- Non-additive errors, non-linear $x \rightarrow y$ effects
- Binary or other non-normal outcomes or exposures
- Differential errors ($Cov(y_j, u_j) \neq 0$), or dependent errors ($Cov(\varepsilon_j, u_j) \neq 0$)), or both
- ME in DV or confounders
- Biases *toward* and *away from* the null are both possible $\Longrightarrow$ ME may boost observed predictive power of aggregate formative scores
- Good news: measurement errors in surveys and regression models are quite well-studied
- Easy to find an appropriate way to either control for ME directly or assess the robustness of key findings to most probable ME scenarios

More good news

- Most common types of ME in survey data are well-known:
  (1) response style
  (2) straightligning
  (3) systematic non-response
  (4) contradicting responses
  (5) duplicated observations
  (6) country-specific translation errors
  (7) other data provider's errors.
- Country scores on (1-5) can easily be computed using survey data and adjusted for in inferential analyses
- As to (6-7), country-specific weird scores quickly become visible to investigators and can simply be excluded from inferential models
  - Change in the attitude toward military rule in Vietnam, Albania and Iran between WVS-3 and WVS-4: Kurzman 2014 in *Monkey Cage*)

Brief illustration with ANDs

- Here ANDs are y, not x, but consequences of $u_{j,y}$ are similar: unbiased β_x but larger residual variance and inflated SEs.
- ANDs as attitudinal imprints of historical legacies
- Only "Theocracy" and "Army" ("Obedience" is more conceptually ambigious)
- H1: In religious countries, the theocratic misunderstanding should be more prevalent
- H2: In more violent environments, the military misunderstanding should be more prevalent
- H3: Measurement error indicators should have less impact on ANDs than substantive indicators
  - Four ME measures: national shares of affirmative, contradictory, duplicated and missing responses inWVS-6 (data from Kirsch and Welzel 2019)


Antecedents of AND national means
DV
Theocracy
Army
ANDs: 2 items
Share of Muslims
Av. religiosity
Repression score
Choice score
LND score
Non-responses
Contradictory
Affirmative
Duplicate
-0.5
0.0
0.5
1.0
Standardized estimate

Summary

- The strongest correlation for "Theocracy" is that with the average religiosity (alone explains 53.4% of variance), while for "Army" it is with the repression score (63.5%);
- Correlations of both items with theoretically relevant variables are generally stronger than their correlations with measurement error indicators;
- Among ME indicators only two, the proportion of contradictory responses and the affirmation rate are significantly (but not very strongly) related to AND items.
- Cross-national variation in the means of two AND items reflects substantive macro-level processes to a larger extent than ME.
- Nomological nets of AND items are not perfectly identical, thus justifying their amalgamation into a single summary score

Practical recommendations

- Two sources of formative comparability:
  - Normative benchmarks
  - Predictive validity: strong correlations with validated and theoretically meaningful country-level measures
- Formative measures' major drawback is that they may be very noisy:
  - Exclude most suspicious (*outlying*) countries
  - Control for survey-based indicators of ME in validation and inferential analyses (should exhibit much smaller effects than substantive correlates)
  - Create an explicit ME model (graphical or formal) and test the sensitivity of your key inferences to different amounts and scenarios of ME.

Take-home message

- Measurement democracy:
  - Reflective measurement is not the single option
  - Formative measures are also in the race
- No omnipotent and universally applicable methods exist:
  - You may not trust them just because they *seem (or promise) to fit* your specific purpose or desire better $\rightarrow$ need for measurement *checks and balances*
- Accuracy of numerical estimates (R) vs. accuracy of effect direction (R/F)

Thank you very much for your attention!

Appendix I: Further reading

Measurement invariance: Basics and PS applications

- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. Annual review of sociology, 40.
- Ariely, G., & Davidov, E. (2011). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Value Survey. Social Indicators Research, 104(2), 271-286.
- Alemán, J., & Woods, D. (2016). Value orientations from the world values survey: How comparable are they cross-nationally?. Comparative Political Studies, 49(8), 1039-1067.
- Sokolov, B. (2018). The index of emancipative values: Measurement model misspecifications. American Political Science Review, 112(2), 395-408.

Basics of formative measurement

- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of consumer research, 30(2), 199-218.
- Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. Journal of Business Research, 61(12), 1219-1228.
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. MIS quarterly, 293-334.
- Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. Psychological methods, 22(3), 581.

Comparability of formative measures

- Diamantopoulos, A., & Papadopoulos, N. (2010). Assessing the cross-national invariance of formative measures: Guidelines for international business researchers. Journal of international business studies, 41(2), 360-370.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. International Marketing Review, 33(3), 405-431.

Formative measurement in PS comparative surveys

- Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. Comparative Political Studies, 49(8), 1068-1094.
- Kirsch, H., & Welzel, C. (2019). Democracy misunderstood: authoritarian notions of democracy around the globe. Social Forces, 98(1), 59-92.
- Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An Overstated Problem With Misconceived Causes. Sociological Methods & Research, 004912412199552 1.

Appendix II: Additional details on the measurement validity analysis of ANDs

Predicting country means and variances of "Theocracy"

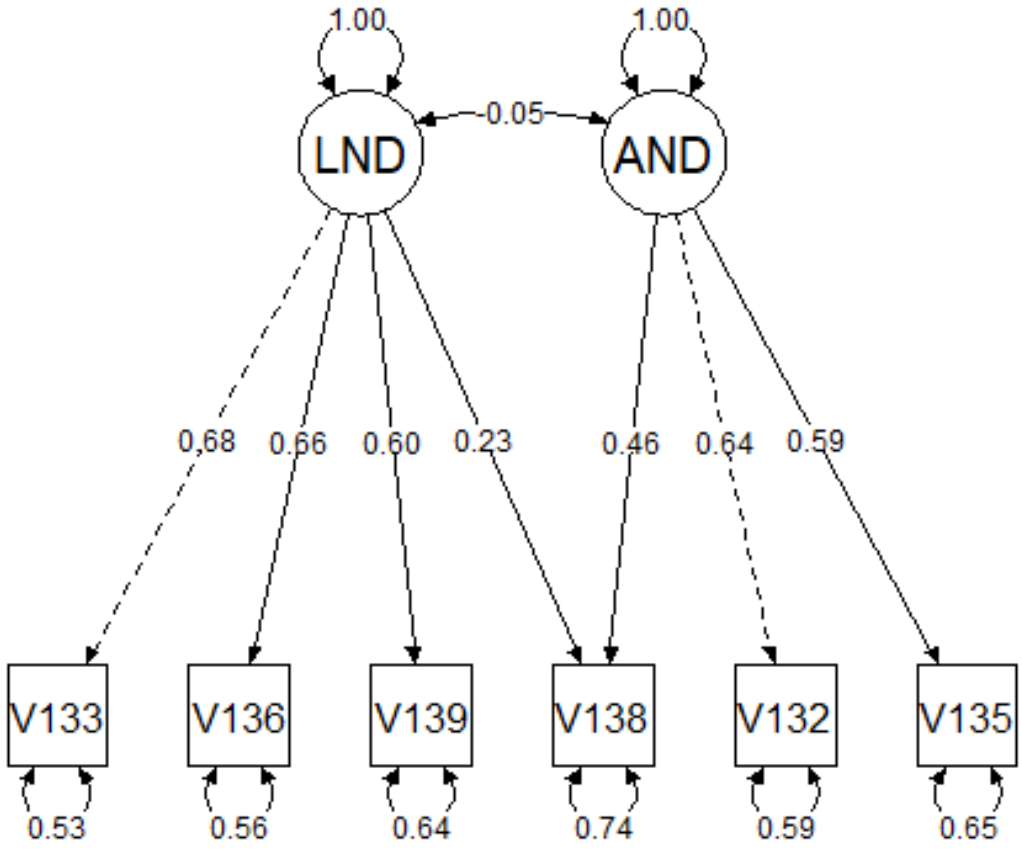
- Share of Muslims in country (from WVS) explains 26.3% of variance in means and 24.2% of variance in variances (both values are adj. R^2s from a bivariate regression)
- Country-mean religiosity score (WVS items about belief, practice and belonging) explains 53.4% and 37.1%, respectively
- For affirmative response style (measured using the science-vs.-religion battery), the respective figures are 25.7% and 19.8%
- For the proportion of contradictory response (measured using questions about interest in politics), 9.2% and 16.3%.

Predicting country means and variances of "Army"

- Repression score (see Kirsch and Welzel 2019) explains 63.5% of variance in means and 22.3% of variance in variances (both values are adj. R^2s from a bivariate regression)
- For affirmative response style (measured using the science-vs.-religion battery), the respective figures are 20% and 11.6%
- For the proportion of contradictory response (measured using questions about interest in politics), 22.4% and 12.1%.
- **Any ideas about appropriate measures of exposure to military rule?**

Appendix III: Reflective analyses of LNDs and ANDs

CFA model of LNDs and ANDs: pooled data



CFA model of LNDs and ANDs: pooled data

- Model fit (MLR-based robust statistics): $\chi^2 = 1377.442$ (df $= 7$, p $= 0.000$), CFI $= 0.976$, TLI $= 0.948$, RMSEA $=$ 0.050(0.048-0.052), SRMR $= 0.018$
- Acceptable globally, better than the model with simple structure (without the cross-loading between LNDs and the obedience item)
- Weak to moderate convergent validity ($\alpha = 0.54$, $\omega = 0.66$, AVE $= 0.37$ only), weak discriminant validity (due to the cross-loading)
- V138 (obedience) is a problematic indicator, which seems logical: "obedience to rulers" is a quite general and vague concept
- Configural invariance:
  - Plausible for LNDs: in all countries loadings are positive, relatively large, and mostly balanced.
  - Does not hold for ANDs: loadings are highly unbalanced, some are negative in some countries.

Standard MGCFA invariance tests of LNDs across 60 WVS-6 countries

Model	χ^2	df	CFI	RMSEA (90% CI)	SRMR
Configural	0.000	0	1.000	.000(.000 " .000)	.000
Metric	750.918	118	.980	.072(.067 " .077)	.032
Partial metric (V139 free)	232.149	59	.995	.053(.046 " .060)	.015
Partial metric (V136 free)	356.863	59	.990	.071(.064 " .078)	.020
Partial metric/scalar (V139 free)	2944.539	118	.926	.140(.136 " .145)	.046

Summary

- (At least partial) metric invariance holds
  - $\rightarrow$ one can compare covariances and regression coefficients for LNDs across countries
- (Partial) scalar invariance does not hold
- Latent mean scores are not comparable?

Flexible novel approaches to invariance testing

- Bayesian approximate approach
  - Allows group-specific measurement parameters to deviate from their sample average estimates
  - The permitted amount of non-invariance is controlled by setting small prior variances on distributions of group-specific deviations
- Alignment
  - Conceptually similar to rotation in exploratory factor analysis
  - Estimate factor means and variances freely in order to minimize non-invariance in the data
- Multilevel CFA (not used here, but related issues are discussed further in this presentation)

Bayesian AMI results

Prior variance	Npar	BIC	DIC	pD	PPP	95% CI
0.001	540.00	661436.47	656036.72	371.46	0.00	2964.72 - 3273.98
0.005	540.00	658869.66	653609.50	441.29	0.00	587.84 - 823.09
0.01	540.00	658446.11	653238.69	467.87	0.00	218.74 - 425.60
0.05	540.00	658122.26	652999.65	510.39	0.16	-42.98 - 142.28
0.1	540.00	658082.37	652980.56	520.72	0.32	-70.08 - 113.75

Bayesian AMI results

- Model with prior variance of group deviations of 0.05 has reasoable fit measures (PPP > 0.05)
- But increasing the prior variance of group-specific intercepts and loadings up to 0.1 improves model fit even further(but perhapsnot much: $\delta_{DIC} = 19.1$)
- Prior variance of 0.05 may be too large to allow for a precise recovery of latent mean scores (e.g. Pokropek, Davidov and Schmidt 2020).
- Many significantly non-invariant parameter estimates.
- Partially invariant models show basically the same fit.
- Approximate scalar invariance doesn't seem to be a plausible assumption

Fixed alignment fit statistics

Items	Loadings			Intercepts		
	Fit Function Contribution	R2	%(N) of non-invar, groups	Fit Function Contribution	R2	%(N) of non-invar. groups
Free elections	-685.590	0.782	5%(3)	-858.549	0.797	38.3%(23)
Civil rights	-793.438	0.297	11.7%(7)	-812.691	0.757	55%(33)
Gender equality	-786.699	0.345	21.7%(13)	-1077.095	0.599	36.7%(22)

Note: All 60 countries from the 6th WVS wave were used. The reference country was South Africa (country code = 47). **Average Invariance index = 0.596**.

- Small proportions of non-invariant groups for loadings (good) but large (all $>$ 25%) for intercepts (not good)
- Still relatively large R^2 for intercepts (and moderate but pluasible R^2 for the whole model). Not so for loadings but see next slide
- V139 seems to be the least reliable item

Alignment simulations and summary

- Use parameters from an estimated alignment model to check how well the model would perform if those were true population values
- Over 500 replications and for realistic group sample sizes (1000, 1500, and 2000 observations per group), correlations between true group means and estimated means are all $>$ 0.99, which is (perhaps) good according to Asparouhov and Muthen (2014).
- Relative biases for particular groups are small, 95% CI coverages are also relatively good
- Simulations revealed some problems with latent variance estimates, but those turned out to be produced by only two problematic countries: Haiti and Kuwait (recall low $R^2 s$ for V136 and V139 loadings on the previous slide)
- Alignment is able to recover latent means and variances with reasonable precision.

Correlations between LND means produced by different approaches