

Comparative Research is Harder Than We Thought: Regional Differences in Experts' Understanding of Electoral Integrity Questions

Bruno Castanho Silva¹ and Levente Littvay²

¹ Cologne Center for Comparative Politics, University of Cologne, Herbert-Lewin Str. 2, D-50931, Cologne, Germany.
Email: bcsilva@wiso.uni-koeln.de

² Department of Political Science, Central European University, Nador u. 9, 1051 Budapest, Hungary

Abstract

Expert evaluations about countries form the backbone of comparative political research. It is reasonable to assume that such respondents, no matter the region they specialize in, will have a comparable understanding of the phenomena tapped by expert surveys. This is necessary to get results that can be compared across countries, which is the fundamental goal of these measurement activities. We empirically test this assumption using measurement invariance techniques which have not been applied to expert surveys before. Used most often to test the cross-cultural validity and translation effects of public opinion scales, the measurement invariance tests evaluate the comparability of scale items across any groups. We apply them to the Perceptions of Electoral Integrity (PEI) dataset. Our findings suggest that cross-regional comparability fails for all eleven dimensions identified in PEI. Results indicate which items remain comparable, at least across most regions, and point to the need of more rigorous procedures to develop expert survey questions.

Keywords: measurement invariance, expert surveys, survey design, electoral integrity

A large part of public opinion research is dedicated to studying whether individuals in different groups understand survey questions in similar ways (see Davidov, Muthén, and Schmidt 2018, for an overview). However, while those using expert surveys have been concerned for a while with systematic differences in experts' response patterns (e.g., Bakker *et al.* 2014), few studies to date applied measurement invariance (MI) techniques from public opinion research to expert surveys (for a recent exception, see Marquardt and Pemstein 2018). Due to respondents' homogeneity and usage of anchoring vignettes, one expects these problems to be minor or absent. Nevertheless, since measuring and comparing country averages is their main goal, if specialists on different countries or regions answer questions differently, comparison becomes impossible.

We show how MI techniques from mass surveys can be applied to analyze differing response patterns from experts using the Perceptions of Electoral Integrity dataset (PEI, Norris, Wynter, and Grömping 2017). Results indicate that direct mean comparisons of countries' electoral integrity across regions are biased across most integrity's dimensions. We conclude with recommendations for those using this dataset and for the elaboration of better expert surveys in general.

1 Data

The PEI dataset¹ asks experts, since 2013, to evaluate whether national elections all around the world can be considered free and fair.² It has 49 questions capturing eleven dimensions of electoral integrity. With the exception of North Africa ($n = 41$), all regions have at least 100

Authors' note: We owe a debt of gratitude to the PEI team, especially Thomas Wynter and Kristina Gushchina who put us in contact. Wynter and his team's responsiveness to our inquiries were exemplary, he explicitly encouraged us to research language effects and worked with us to get what we needed. We are also thankful for comments on earlier versions of this manuscript by the three anonymous reviewers and Jeff Gill. The replication files for this article are available on the Political Analysis Dataverse (Castanho Silva and Littvay 2019).

1 Version 5.5, October 2017.

2 More details on sampling and methods in the Online Appendix.

Political Analysis (2019)
vol. 27:599–604
DOI: 10.1017/pan.2019.24

Published
3 July 2019

Corresponding author
Bruno Castanho Silva

Edited by
Jeff Gill

© The Author(s) 2019. Published
by Cambridge University Press
on behalf of the Society for
Political Methodology.

expert respondents.³ The dimensions of electoral integrity are: 1. Electoral Laws (3 questions); 2. Electoral Procedures (4); 3. Boundaries for Voting Districts (3); 4. Voter Registration (3); 5. Party and Candidate Registration (5); 6. Media Coverage (5); 7. Campaign Finance (5); 8. Voting Process (8); 9. Vote Count (5); 10. Voting Results (4); 11. Election Authorities (4). Responses have a 1–5 Likert scale, from “Strongly disagree” to “Strongly agree.” For example, on “Election Authorities,” the first indicator reads “The election authorities were impartial.”⁴

2 Measurement Invariance

With Likert scales, MI is obtained if two respondents who have the same level of a latent construct give the same answer to a question measuring it, regardless of group membership (Millsap 2011).⁵ We test two types of invariance: metric and scalar (Millsap 2011). Metric invariance means that an increase of 1 unit in the latent trait corresponds to the exact same increase in the response for any individual, regardless of group identity. It yields comparable coefficients when used in regression analysis. Scalar invariance requires, on top of metric, that two individuals with the same level of the construct would give the exact same answer. Scalar invariance is more demanding, but necessary for comparison of group means.

We test MI using multigroup confirmatory factor analysis (MGCFAs, Jöreskog 1971). In a CFA model, indicators theorized to measure each dimension should load together into a single latent variable. MGCFAs consider that observations might be nested into qualitatively different groups. It allows researchers to test if some estimated model parameters differ across the groups: for example, a factor loading may have one value for respondents from region A and another for B, while other model parameters are estimated as the same across all regions. If the measurement instrument is invariant across groups, freely estimated loadings should be similar: if the loading of an indicator is high for one region but low for another, it suggests that the items function differently. We apply MGCFAs with three nested invariance models: configural, metric (factor loadings are constrained to be equal across groups) and scalar (loadings and intercepts are constrained to be equal across groups), testing invariance by comparing the χ^2 -distributed difference in the $-2 \times \log$ -likelihoods.

The χ^2 invariance test in MGCFAs has been criticized for finding noninvariance even in the presence of substantively minor deviations (Oberski 2014). Therefore, we also present the Root Mean Squared Error of Approximation (RMSEA), and the Comparative Fit Index (CFI). These are less sensitive to large samples and indicate whether the more restricted models have acceptable fit. Next we use the Alignment method (Asparouhov and Muthén 2014). This is an optimization approach to invariance which minimizes overall noninvariance of the model while allowing noninvariance of a few indicators or groups (Asparouhov and Muthén 2014). While these may increase the odds of Type II errors (not finding noninvariance that exists), combining both methods gives us a good overview of actual (non)invariance levels.⁶

We test MI in each of the eleven electoral integrity dimensions, each modeled as a latent variable with three to eight indicators as denoted in the codebook. Invariance is tested across two clustering options: five continents, or nine regions (defined by PEI itself). Instead of using the

3 Descriptive statistics in the Online Appendix, Table S1.

4 All items are in Table 2. Replication data and codes for all analysis in this paper are available on the Political Analysis Dataverse (Castanho Silva and Littvay 2019).

5 Measurement (non)invariance is conceptually the same as Differential Item Functioning (DIF, Kline 2016, p. 398). The main difference is that MI assumes a linear relationship between latent construct and indicators, treating indicators as continuous. DIF assumes a nonlinear relationship, treating variables as ordinal—for empirical and mathematical comparisons, see Muthén and Christofferson (1981), and Raju, Laffitte, and Byrne (2002). The PEI aggregation assumes continuous measurement, by summing up and standardizing responses (Norris, Wynter, and Grömping 2017), in common with most expert surveys, so we apply Structural Equation Modeling based MI tests.

6 For a deeper discussion of advantages of different invariance tests, see Davidov, Muthén, and Schmidt (2018).

Table 1. MGCFA test of measurement invariance.

Dimension	$\Delta\chi^2(df)$	<i>p</i> -value	RMSEA (Δ)	CFI (Δ)
5 Continents				
Laws	10.274(8)	0.246	0.025(0.025)	0.999(0.001)
Procedures	20.737(12)	0.054	0.053(0.003)	0.987(0.006)
Boundaries	6.509(8)	0.590	0.000(0.000)	1.00(0.000)
Voter registration	6.407(8)	0.602	0.000(0.000)	1.00(0.000)
Party registration	83.399(16)	<i>p</i> < 0.001	0.222(.031)	0.752(0.051)
Media coverage	35.102(16)	0.004	0.169(0.027)	0.871(0.023)
Campaign finance	16.211(16)	0.438	0.184(0.032)	0.835(0.026)
Voting process	69.626(28)	<i>p</i> < 0.001	0.135(0.008)	0.874(0.015)
Vote count	29.051(16)	0.024	0.217(0.016)	0.657(0.102)
Voting results	27.646(12)	0.001	0.058(0.003)	0.987(0.008)
Electoral authorities	13.359(12)	0.343	0.064(0.011)	0.982(0.007)
9 Regions				
Laws	39.141(16)	0.001	0.077(0.077)	0.987(0.013)
Procedures	47.254(24)	0.003	0.061(0.000)	0.985(0.009)
Boundaries	12.630(16)	0.699	0.00(0.000)	1.00(0.000)
Voter registration	20.131(16)	0.214	0.037(0.037)	0.996(0.004)
Party registration	76.600(32)	<i>p</i> < 0.001	0.224(0.048)	0.778(0.030)
Media coverage	55.279(32)	0.006	0.180(0.043)	0.849(0.015)
Campaign finance	53.01(32)	0.011	0.205(0.045)	0.816(0.024)
Voting process	121.10(56)	<i>p</i> < 0.001	0.166(0.016)	0.838(0.014)
Vote count	386.73(32)	<i>p</i> < 0.001	0.268(0.005)	0.595(0.159)
Voting results	33.036(24)	0.103	0.062(0.017)	0.985(0.005)
Electoral authorities	23.090(24)	0.514	0.077(0.022)	0.974(0.007)

$\Delta\chi^2$: difference between the configural and metric invariance models' χ^2 statistics; *df*: difference in the number of free parameters between the two models, and *p*-values for this difference. RMSEA and CFI from the metric invariance models; Δ RMSEA and Δ CFI are the difference in RMSEA and CFI from that model to the configural. All fit statistics are calculated with the Satorra and Bentler (2001) correction for robust estimation.

original responses, we run the analyses on DIF-corrected rank ordered responses obtained with PEI's anchoring vignettes, following the nonparametric model by King *et al.* (2004).⁷

3 Results

Table 1 has results of metric invariance tests with MGCFA. We do not report scalar invariance because no dimension even approaches it.⁸ For no dimensions of electoral integrity can we directly compare mean values of countries across regions. A few dimensions fare well in metric invariance: "Boundaries," "Voter Registration," and "Electoral authorities," in both regions and continents. Most others not only display significant differences between the configural and metric models, but the models with constrained loadings have very poor fit, as denoted by RMSEAs greater than 0.05 and CFIs less than 0.95 (recommended thresholds for good fit by Hu and Bentler 1999). It is important to highlight that noninvariance remains after incorporating anchoring vignettes into the model,⁹ confirming the finding by von Davier *et al.* (2017) that vignette questions

7 Vignettes changed in 2017, so we only use pre-2017 data in these tests. Technical details of the vignettes and their incorporation, and analyses without vignettes, in the Online Appendix.

8 Results with scalar invariance in the Online Appendix.

9 In the Online Appendix we show other ways beyond DIF-corrected indicators to incorporate vignettes. None produces more invariant results.

Table 2. Alignment method: loadings and intercepts invariance.

Indicator	5 continents					9 regions								
	1	2	3	4	5	1	2	3	4	5	6	7	8	9
Laws	Electoral laws were unfair to smaller parties	0.06												
	Electoral laws favored the governing party or parties													
	Elections laws restricted citizens' rights													
Proced.	Elections were well managed			-0.04										
	Information about voting procedures was widely available			0.08										
	Elections were conducted in accordance with the law													
Bound.	Boundaries discriminated against some parties			-0.09										
	Boundaries favored incumbents				0.04									
	Boundaries were impartial				-0.21									
V. reg.	Some citizens were not listed in the register													
	The electoral register was inaccurate													
	Some ineligible electors were registered													
Party reg.	Some opposition candidates were prevented from running													
	Women had equal opportunities to run for office													
	Ethnic and national minorities had equal opportunities to run for office													
	Only top party leaders selected candidates													
	Some parties/candidates were restricted from holding campaign rallies	-0.17		-0.17										
Coverage	Newspapers provided balanced election news													
	TV news favored the governing party		0.18											
	Parties/candidates had fair access to political broadcasts and advertising													
	Journalists provided fair coverage of the elections													
	Social media were used to expose electoral fraud													
Resources	Parties/candidates had equitable access to public political subsidies	0.09		0.11										
	Parties/candidates had equitable access to political donations													
	Parties/candidates publish transparent financial accounts													
	Rich people buy elections													
	Some state resources were improperly used for campaigning													
Voting process	Some voters were threatened with violence at the polls													
	Some fraudulent votes were cast	-0.42												
	The process of voting was easy													
	Voters were offered a genuine choice at the ballot box													
	Postal ballots were available													
	Special voting facilities were available for the disabled													
	National citizens living abroad could vote													
Some form of internet voting was available														
Count	Ballot boxes were secure													
	The results were announced without undue delay													
	Votes were counted fairly													
	Int'l election monitors were restricted													
	Domestic election monitors were restricted													
Results	Parties/candidates challenged the results													
	The election led to peaceful protests													
	The election triggered violent protests													
	Any disputes were resolved through legal channels	0.30		0.20										
El. Auth.	The election authorities were impartial													
	The authorities distributed information to citizens													
	The authorities allowed public scrutiny of their performance													
	The election authorities performed well													

Numbers indicate the standardized difference between the noninvariant group intercept and the weighed average intercept between invariant groups. Continents: 1: Africa; 2: Americas; 3: Asia; 4: Europe; 5: Oceania. Regions: 1: East & Southern Africa; 2: West & Central Africa; 3: East Asia & Pacific; 4: South Asia; 5: Eastern Europe; 6: Western Europe; 7: Middle East; 8: North Africa; 9: Americas.

themselves can be understood differently by respondents in different groups, ultimately just moving the noninvariance problem one step further.

With the alignment analysis (Table 2), we find that noninvariance in some dimensions is restricted to a few items, and mostly concentrated in some regions. For example, the first indicator for “Electoral Laws” has a noninvariant intercept in Africa, while the third has noninvariant loading and intercept in Western Europe. These indicators, however, work well in all other continents or regions to measure the integrity of electoral laws in a cross-regionally comparable way. In these cases, users of PEI could drop problematic regions or items from their analyses. The best performing dimension is “Voter registration,” where all items have good cross-regional validity and can be used without problems in comparative research. Western Europe is the most problematic region overall, with the highest number of indicators that are noninvariant for loadings and intercepts. We may hypothesize that specialists in Western Europe have different anchors for inappropriate electoral conduct, resulting in different response styles even after accounting for DIF with anchoring vignettes.

Noninvariance of indicators' intercepts is considered the least problematic, since if loadings are invariant one can still get unbiased regression estimates. To check the magnitude of the (smallest possible) problem we look at the size of noninvariance for intercepts. Numbers in the light gray cells are the standardized difference between that indicator's intercept for that group and the weighed average intercept for invariant groups. Looking again at the noninvariant intercept for the third indicator in "Boundaries" in Europe, the estimate means that, for an election with the exact same level of integrity of boundaries for voting districts, an European respondent would give an answer 0.21 standard deviations below the average of respondents from other regions. There are larger differences in other dimensions and regions: for example, experts in Africa evaluating whether "fraudulent votes were cast" ("Voting Process") would give answers 0.42 SDs below the average of other respondents evaluating exactly the same elections. Considering how these questions are aggregated, biases add up to substantively different evaluations of elections across regions.

4 Conclusion

In this letter we perform MI tests in the PEI dataset (Norris, Wynter, and Grömping 2017), and show that experts have different response styles depending on the region where their countries of specialty are located. These results reinforce findings that one should not use aggregate means of expert surveys (Marquardt and Pemstein 2018). The problem persists after using anchoring vignettes, which are supposed to account for DIF. Consequently, it is unlikely that the issue could be solved by increasing the number of experts, or perhaps crowdsourcing (Maestas, Buttice, and Stone 2014): if all specialists on Western Europe or West European citizens have a systematically different understanding of electoral integrity than experts of other regions, errors are not random and will not cancel out by increasing the pool size. Due to this systematic bias on answers to each individual question, using other measures of central tendency (as suggested by Lindstädt, Proksch, and Slapin 2018) would still yield incomparable estimates across regions. Beyond incomparable averages, lack of loadings' invariance also compromises these dimensions' use in regression analysis. For instance, the solutions proposed by Marquardt and Pemstein (2018) can help correcting for scalar invariance and yield cross-regionally comparable country means. They might not, however, solve the lack of metric invariance for linear models and the biased regression estimates that follow. The only solution, given available data, would be to use Bayesian multilevel structural equation models with random factor loadings (Muthén and Asparouhov 2018; Castanho Silva, Bosancianu, and Littvay 2019). Such methods can model the cross-regional noninvariance and should be applied by researchers who use the electoral integrity dimensions in their regression analyses.

There might be a few ways to alleviate noninvariance problems when designing expert surveys. First, following PEI's exemplary lead and having (a) constructs with a strong theoretical foundation, (b) multiple questions per construct, and (c) anchoring vignettes. This allows, at least, to test whether the measurement is cross-regionally valid. Second, one could also collect pilot data with quasi-experts (e.g., internationally diverse groups of social science graduate students) who are likely to respond similarly to the targeted respondents. Following the analytical steps presented here, one can identify, adjust or eliminate problematic items, and, if needed, even drop entire regions. We call for better design of expert survey batteries and more rigorous pretests, leading to comparative measurement of relevant concepts where results can actually be compared.

Supplementary material

For supplementary material accompanying this paper, please visit <https://doi.org/10.1017/pan.2019.24>.

References

- Asparouhov, T., and B. Muthén. 2014. "Multiple-group Factor Analysis Alignment." *Structural Equation Modeling: A Multidisciplinary Journal* 21(4):495–508.
- Bakker, R., E. Edwards, S. Jolly, J. Polk, J. Rovny, and M. Steenbergen. 2014. "Anchoring the Experts: Using Vignettes to Compare Party Ideology Across Countries." *Research & Politics* 1(3):1–8.
- Castanho Silva, B., C. M. Bosancianu, and L. Littvay. 2019. *Multilevel Structural Equation Modeling*. Thousand Oaks, CA: Sage.
- Castanho Silva, B., and L. Levente. 2019. "Replication Data for: Comparative Research is Harder than We Thought: Regional Differences in Experts' Understanding of Electoral Integrity Questions." <https://doi.org/10.7910/DVN/HG2CAJ>, Harvard Dataverse, V1.
- Davidov, E., B. Muthén, and P. Schmidt. 2018. "Measurement Invariance in Cross-National Studies: Challenging Traditional Approaches and Evaluating New Ones." *Sociological Methods & Research* 47(4):631–636.
- von Davier, M., H.-J. Shin, L. Khorramdel, and L. Stankov. 2017. "The Effects of Vignette Scoring on Reliability and Validity of Self-Reports." *Applied Psychological Measurement* 42(4):291–306.
- Hu, L.-t., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6:1–55.
- Jöreskog, K. G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 36:409–426.
- King, G., C. J. L. Murray, J. A. Salomon, and A. Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1):191–207.
- Kline, R. B. 2016. *Principles and Practice of Structural Equation Modeling*. 4th edn. New York: The Guilford Press.
- Lindstädt, R., S.-O. Proksch, and J. B. Slapin. 2018. "When Experts Disagree: Response Aggregation and its Consequences in Expert Surveys." *Political Science Research and Methods*, <https://doi.org/10.1017/psrm.2018.52>.
- Maestas, C. D., M. K. Buttice, and W. J. Stone. 2014. "Extracting Wisdom from Experts and Small Crowds: Strategies for Improving Informant-based Measures of Political Concepts." *Political Analysis* 22(3):354–373.
- Marquardt, K. L., and D. Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26(4):431–456.
- Millsap, R. E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Routledge.
- Muthén, B. O., and A. Christofferson. 1981. "Simultaneous Factor Analysis of Dichotomous Variables in Several Groups." *Psychometrika* 46(4):407–419.
- Muthén, B. O., and T. Asparouhov. 2018. "Recent Methods for the Study of Measurement Invariance With Many Groups: Alignment and Random Effects." *Sociological Methods & Research* 47(4):637–664.
- Norris, P., T. Wynter, and M. Grömping. 2017. *Codebook – The Expert Survey of Perceptions of Electoral Integrity, Release 5.5, (PEI_5.5)*. The University of Sidney: The Electoral Integrity Project.
- Oberski, D. L. 2014. "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models." *Political Analysis* 22(1):45–60.
- Raju, N. S., L. J. Laffitte, and B. M. Byrne. 2002. "Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory." *Journal of Applied Psychology* 87(3):517–529.
- Satorra, A., and P. M. Bentler. 2001. "A Scaled Difference chi-Square Test Statistic for Moment Structure Analysis." *Psychometrika* 66(4):507–514.