Erschienen in: Politics and Governance ; 6 (2018), 1. - S. 92-104
https://dx.doi.org/10.17645/pag.v6i1.1235

COGITATIO

Politics and Governance (ISSN: 2183–2463)
2018, Volume 6, Issue 1, Pages 92–104
DOI: 10.17645/pag.v6i1.1235

Article

Method Factors in Democracy Indicators

Martin Elff [1] and Sebastian Ziaja [2,*]

[1] Faculty of Political & Social Sciences, Zeppelin University, 88045 Friedrichshafen, Germany; E-Mail: martin.elff@zu.de
[2] Research Center for Distributional Conflict and Globalization, Heidelberg University, 69115 Heidelberg, Germany; E-Mail: ziaja@uni-heidelberg.de

* Corresponding author

Submitted: 20 October 2017 | Accepted: 22 December 2017 | Published: 19 March 2018

Abstract
Method factors represent variance common to indicators from the same data source. Detecting method factors can help uncover systematic bias in data sources. This article employs confirmatory factor analysis (CFA) to detect method factors in 23 democracy indicators from four popular data sources: The Economist Intelligence Unit (EIU), Freedom House, Polity IV, and the Varieties of Democracy (V-Dem) project. Using three different multi-dimensional concepts of democracy as starting points, we find strong evidence for method factors in all sources. Method-specific factors are strongest when yearly changes in the scores are assessed. The sources find it easier to agree on long-term average scores. We discuss the implications for applied researchers.

Keywords
confirmatory factor analysis; democracy; democracy indicators; measurement; method bias

Issue
This article is part of the issue "Why Choice Matters: Revisiting and Comparing Measures of Democracy", edited by Heiko Giebler (WZB Berlin Social Science Center, Germany), Saskia P. Ruth (German Institute of Global and Area Studies, Germany), and Dag Tanneberg (University of Potsdam, Germany).

1. Introduction

Are measures of political regime characteristics systematically influenced (or biased) by the institutions that created them? We answer this question by assessing whether democracy indicators coming from the same source exhibit common deviations not found in indicators originating from other sources. Kenneth Bollen (1993) referred to these common deviations as 'method factors'. That is, we assess whether democracy indicators are affected by method factors.

Method factors should be of concern for the applied researcher because method factors can be a sign of systematic bias in the data source, i.e., 'method bias'. But biased indicators do not just lead to improper descriptions of the issue to be measured. Even erroneous conclusions can be drawn from inferential studies if bias in a democracy indicator is correlated with an explanatory variable such as economic liberalization or ethnic fragmentation. In a field with immediate policy implications such as democratization research, false conclusions can cause actual harm. There is peril in informing policy with biased indicators.

This decade has seen a renewed interest in measuring democracy and explaining why it succeeds in some places and fails in others. The interest was sparked by initially promising signs of liberalization in the Middle East, now known as the Arab spring, as well as by formally democratized countries backsliding into autocratic practice. Such backsliding has occurred prominently in Russia, Turkey and Venezuela, but also in member states of the European Union, such as Poland and Hungary.

Responding to the desire to better track these developments, several new producers of democracy data have come forward. Certainly, the most notable addition to the group of democracy data producers is the Varieties

of Democracy (V-Dem) project. It boasts thousands of experts who have been involved in coding a new dataset on democracy for almost all countries since 1900. 'Most [experts] have lived in their countries of expertise for nearly thirty years, and at least 60 percent are nationals of that country' (Lindberg, Coppedge, Gerring, & Teorell, 2014, p. 162). The size and diversity of V-Dem's expert group contrasts with existing data sources, such as the Polity IV project or the FH data, which rely on a much smaller number of experts who are predominantly citizens of the United States. This difference in expert groups raises the question whether V-Dem data differs systematically from existing data sources.

The suspicion that traits of the expert group could influence indicators has already been voiced by Bollen (1993, p. 1213). He lists political attitudes of coders, incomplete information, and the aggregation of individual indicators into indices of democracy as potential sources of method bias. All of these issues are present to varying degrees in all social science indicators. One might assess these issues from varying perspectives, examining underlying concepts, data collection and aggregation procedures.

In this article, we focus on the indicator scores that the sources produce. We employ a 'convergent/divergent validation' approach (Adcock & Collier, 2001, p. 540): indicators from one source representing a particular conceptual dimension of democracy should be similar to indicators from other sources representing the same dimension—they should converge. Indicators from the same source representing different conceptual dimensions of democracy should not be as similar—they should diverge to some degree.

Our specific strategy is inspired by Bollen's seminal 1993 article 'Liberal Democracy: Validity and Method Factors in Cross-National Measures'. He employs confirmatory factor analysis (CFA). His model allows a range of indicators to load on two latent dimensions of democracy—*political liberties* and *democratic rule*—and on latent factors representing the sources of the indicators. This enables him to assess the amount of systematic error that indicators from the same source exhibit, i.e., the method factors. The sources he considers are Arthur Banks' *Cross-National Time-Series Data Archive* as well as Freedom House's *Freedom in the World* and *Freedom in the Press*. Bollen's analysis has not been updated with current democracy indices. Giebler (2012, p. 510) argues that most approaches comparing democracy measures focus excessively on conceptual differences, rather than on methodological differences. A study that has paid much attention to systemic bias in democracy ratings in the past decade is Pemstein, Meserve and Melton's (2010, pp. 444–446) presentation of their *Unified Democracy Scores*. What distinguishes their approach from ours is that they employ top-level indices that attempt to capture democracy as a whole. Also, Treier and Jackman (2008), in an attempt to detect measurement error in the Polity data, employ a unidimensional model.

Our contribution to the literature is an update and extension to Bollen's approach that evaluates various measures of regime characteristics that have been used as indicators of unidimensional or multidimensional conceptualisations of democracy. We update his approach by employing an updated set of four sources and 23 indicators and three different conceptual frameworks in our analysis. We extend his approach by using data with temporal variation over the period 2006 to 2014 (and 1972 to 2014 for an alternative dataset with three sources). Temporal variation provides us with the ability to provide a more nuanced assessment: do sources agree on average on grading countries by traits of democracy and do they agree about the timing and direction of the changes taking place over time?

The main result of our analysis is that most if not all measures of regime characteristics under study exhibit method factors. That is, these measures are influenced to a non-ignorable degree by the sources or institutes that produce them. The objectivity of these measures is thus limited. This is particularly salient when one considers changes over time in democracy indicators: different sources make very different assessments on changes in political and civil rights. In the cross-sectional view, however, sources show less pronounced method factors and more substantial agreement. Considering individual sources, method factors are largest for some of the Polity IV indicators.

2. Background: Three Concepts of Democracy

Our aim is to examine whether current measures of democracy are affected by the institutions or research groups by which they are produced or by the data sources from which they are derived. To ensure that our findings regarding the existence of method factors do not depend on a particular conception of democracy, we check for the existence of method factors against the backdrop of each of three different conceptualisations: a two-dimensional one, which has already been employed in Bollen's (1993) study, a three-dimensional conceptualisation put forward by the V-Dem project (Coppedge, Gerring, Lindberg, Skaaning, & Teorell, 2017a), and finally a conceptualisation with no less than four dimensions, inspired by Merkel (2004).

Bollen's two-dimensional scheme entails *political liberties* and *democratic rule*. This approach is rather minimalist and introduces only one distinction: between individuals' abilities to participate in the political system, and the functioning of the latter in the spirit of democracy. Using this approach has the advantage of making our analysis comparable to Bollen's. The dimensions are defined as follows: *Political liberties* 'exist to the extent that the people of a country have the freedom to express a variety of opinions in any media and the freedom to form or to participate in any political group' (Bollen, 1993, p. 1208). *Democratic rule* '(or political rights) exists to the extent that the national government is ac-

countable to the general population, and each individual is entitled to participate in the government directly or through representatives' (Bollen, 1993, p. 1209).

Since we have more indicators available than Bollen had in 1993, we can test more detailed models. V-Dem suggests a scheme which entails seven 'principles' of democracy: electoral, liberal, participatory, majoritarian, consensual, deliberative and egalitarian (Coppedge et al., 2017a, pp. 20–25). We employ only the first three principles here and exclude the latter four. We exclude majoritarian and consensual, as these principles do not have a more democratic and a less democratic pole (Lijphart, 2012)—rather they refer to variations of democracy which most measurement projects do not tap into. Measuring majoritarian and consensual principles across all countries is challenging, and even V-Dem abstains from quantifying these concepts at the moment (Coppedge et al., 2017a, p. 24). We exclude the deliberative principle because it refers to the rationality of political debates, which is not measured directly by other sources. We exclude the egalitarian principle because it refers to individual socio-economic prerequisites for political empowerment, which includes financial resources and thus seems to overstretch the concept of democracy.

The *electoral principle* of V-Dem captures the core idea of polyarchy, i.e., 'making rulers responsive to citizens through periodic elections' (Coppedge et al., 2017a, p. 20). The *liberal principle* refers to individual rights against state repression, guaranteed by 'constitutionally protected civil liberties, strong rule of law, and effective checks and balances that limit the use of executive power' (Coppedge et al., 2017a, p. 21). The *participatory principle* 'embodies the values of direct rule and active participation by citizens in all political processes' (Coppedge et al., 2017a, p. 22). 'Direct rule' entails problems similar to that encountered in the majoritarian and consensual principles: more direct rule does not necessarily imply more democracy, and may erode into a tyranny of the majority. Moreover, most sources abstain from measuring issues of direct democracy separately. We thus focus on 'active participation' as conceptual focus of this dimension.

A more detailed theoretical model which is independent of our data sources is Wolfgang Merkel's (2004) concept of 'embedded democracy'. Embedded democracy has an *electoral regime* at its core, which is complemented by *political liberties*, *civil rights*, *horizontal accountability*, and the *effective power to govern* (Merkel, 2004, p. 37). Referring to Dahl (1971), Merkel (2004, p. 38) describes the *electoral regime* to entail 'universal, active suffrage, universal, passive right to vote, free and fair elections and elected representatives'. *Political rights* 'complete the vertical dimension of democracy and make the public arena an independent political sphere of action, where organizational and communicative power is developed' (Merkel, 2004, p. 38). This requires freedom of association and freedom of expression. Political rights provide the input for the electoral regime, which would be lacking input without the former. *Civil rights* maintain the rule of law by containing state power. 'The actual core of the liberal rule of law lies in basic constitutional rights' (Merkel, 2004, p. 39). This requires an independent judiciary as a guarantor. Civil rights thus constitute negative rights in the political system, whereas political rights constitute positive rights. Merkel refers to Guillermo O'Donnell (1994) with the definition of *horizontal accountability*, which requires 'that elected authorities are surveyed by a network of relatively autonomous institutions' (Merkel, 2004, p. 40). This is necessary since the vertical forms of control provided by the three preceding institutions 'control the government only periodically'. The *effective power to govern* finally asserts that the elected representatives are actually in control of the state (Merkel, 2004, p. 41). We disregard this last requirement here, as only a few sources attempt to measure it.

3. Data

The data sources we consider beyond V-Dem are the Economist Intelligence Unit's (EIU) democracy index (Economist Intelligence Unit, 2014), Freedom House (FH; Freedom House, 2016) and the Polity IV project (Marshall, Gurr, & Jaggers, 2016). As Bollen (1993, p. 1210) does, we focus on subjective measures, i.e., indicators of *de facto* democratic quality, not *de jure* provisions. The former are also more susceptible to systematic biases and deserve a closer inspection in this regard. Both this focus and the limited coverage of other sources explain why we constrain our analysis to four sources at the present.

Other sources were excluded for providing discrete regime types instead of linear measures of regime status (e.g. the Democracy-and-Dictatorship data by Cheibub, Gandhi, & Vreeland, 2010), for measuring *de jure* instead of *de facto* regime traits (e.g., the Database of Political Institutions by Beck, Clarke, Groff, Keefer, & Walsh, 2001), or for providing insufficient spatial or temporal coverage (e.g., the Bertelsmann Transformation Index, 2016). The Appendix lists additional sources that were excluded and gives reasons for these decisions.

Each data source considered here provides at least two levels of indicators: those at the lowest level, which are coded directly (by judgement, observation or other means), and those at intermediate and higher levels, which are aggregated from lower-level indicators. We employ data at an intermediate level of aggregation, i.e., indicators that are supposed to measure rather general attributes of democratic rule and political liberties, such as fair elections and freedom of speech. Disregarding very detailed indicators such as those provided by V-Dem allows us to maintain a roughly equal level of aggregation across sources, although a perfect alignment is not possible. The selection of the indicators and their assignment to conceptual dimensions is documented at length in the Appendix.

The indicators are collated into two data sets: a 'longer' data set which includes a smaller range of indicators for which data are available for a relatively long period, from 1972 to 2014, and a 'shorter' but 'wider' data set which includes a wider range of indicators for which data are available only for the relatively short period from 2006 to 2014. The 'longer' data set, to which we refer to as *D1*, contains 5,864 observations from 160 countries for 15 indicators, while the 'shorter' data set, referred to as *D2*, contains 1,070 observations from 157 countries for 23 indicators. The EIU's democracy data and detailed indicators are only available for more recent years and appear in *D2*, but not in *D1*.

The following results focus on *D2*. Results pertaining to *D1* confirm the general findings from *D2* and can be found in the Appendix. A description of *D2* is given by Table 1, which indicates the source of the indicators and what dimensions they represent according to conceptualisations of political regimes with different numbers of dimensions. The Appendix also provides tables with summary statistics for both datasets. All data was obtained via the *Quality of Government* database (Teorell et al., 2017) and the V-Dem dataset version 7.1 (Coppedge et al., 2017b).

Table 1. Indicators and dimension assignment in data set *D2*.

Description	Source	2-Dimensional concept	3-Dimensional concept	4-Dimensional concept
Civil liberties	EIU	Political liberties	Liberal principle	Civil rights
Electoral process and pluralism	EIU	Democratic rule	Electoral principle	Electoral regime
Functioning of government	EIU	Democratic rule	Liberal principle	Horizontal accountability
Political participation	EIU	Political liberties	Participatory principle	Political rights
Associational and Organizational Rights	FH	Political liberties	Participatory principle	Political rights
Electoral Process	FH	Democratic rule	Electoral principle	Electoral regime
Freedom of Expression and Belief	FH	Political liberties	Liberal principle	Civil rights
Personal Autonomy and Individual Rights	FH	Political liberties	Liberal principle	Civil rights
Political Pluralism and Participation	FH	Democratic rule	Liberal principle	Political rights
Rule of Law	FH	Political liberties	Liberal principle	Civil rights
The competitiveness of participation (PARCOMP)	Polity	Democratic rule	Participatory principle	Political rights
Regulation of participation (PARREG)	Polity	Political liberties	Liberal principle	Political rights
Executive constraints (XCONST)	Polity	Democratic rule	Liberal principle	Horizontal accountability
Competitiveness of executive recruitment (XRCOMP)	Polity	Democratic rule	Electoral principle	Electoral regime
Openness of executive recruitment (XROPEN)	Polity	Political liberties	Participatory principle	Political rights
Civil society participation	V-Dem	Political liberties	Participatory principle	Political rights
Freedom of association (thick)	V-Dem	Political liberties	Electoral principle	Political rights
Freedom of expression (thick)	V-Dem	Political liberties	Liberal principle	Political rights
Judicial constraints on the executive	V-Dem	Democratic rule	Liberal principle	Horizontal accountability
Equality before the law and individual liberty	V-Dem	Political liberties	Liberal principle	Civil rights
Equal protection index	V-Dem	Political liberties	Liberal principle	Civil rights
Clean elections	V-Dem	Democratic rule	Electoral principle	Electoral regime
Legislative constraints on the executive	V-Dem	Democratic rule	Liberal principle	Horizontal accountability

4. Method

If measures of political regime characteristics are systematically influenced by the institutions that created them, then measures coming from the same institution should be more correlated than those coming from different institutions, at least after taking into account that these measures reflect certain substantive dimensions of regime characteristics. Equivalently, the variation of a measure of regime characteristics can be decomposed into a portion that can be attributed to a variation along conceptual dimensions such as, e.g. *democratic rule* and *political liberties*, and a portion that has to be attributed to the influence of the institution that created the measure. Confirmatory factor analysis (CFA) is the method of choice to examine whether such a decomposition is possible and adequate. In the context of confirmatory factor analysis this decomposition will take the form:

$$X_i = \alpha_i + \lambda_{ij}F_j + \kappa_{ik}G_k + U_i \quad (1)$$

where X_i is the value of the i-th regime characteristics indicator, F_i is the (unobserved) value of a common factor that represents the j-th conceptual dimension, G_k is the (unobserved) value of the common factor that represents the influence of the k-th institution, and U_i is a unique factor that represents random measurement error specific for the i-th indicator. For brevity, in the following we will refer to a common factor that represents a conceptual dimension of regime characteristics as a *conceptual factor*, while we refer to a common factor that represents the influence of the institution that created the indicator as a *method factor*.[1] The coefficient λ_{ij} of the conceptual factor F_j is referred to as the *loading* of indicator X_i on this factor. It is a parameter that is estimated in the context of a confirmatory factor analysis and represents how much the variation of the regime characteristics indicator is influenced by the conceptual factor. The coefficient κ_{ik} of the method factor G_k, which also can be referred to as a loading, is an estimated parameter that represents how much the indicator is influenced by the institute that has created it. Finally, α_j is an intercept that reflects the fact that, while the means of the common factors and the unique factor are assumed to be zero, the mean of X_i may be different from zero.[2]

Figure 1 illustrates a decomposition of the four indicators X_1, X_2, X_3, and X_4, into two conceptual factors F_1 and F_2, and into two method factors G_1 and G_2. The loadings in equation (1) are represented by arrows in Figure 1 as is the influence of the unique factors U_1, U_2, U_3, and U_4, which are represented by the empty circles. Figure 1 also illustrates some additional assumptions that we make in our analysis: firstly, that method factors are uncorrelated and that unique factors are uncorrelated, while conceptual factors may be correlated. These assumptions are motivated by the following considerations: the fact that one can distinguish between concepts such as *Democratic rule* and *Political liberties* does not imply that they are empirically uncorrelated. On the other hand, if the method factors are supposed to reflect influences that are *specific* to the institutions that create the indicators, this is best reflected by assuming the method factors to be uncorrelated. Otherwise, if we allowed the method factors to be correlated, such correlations would reflect commonalities among indicators of different institutions, which in turn could be attributed to the fact that they are supposed to measure the same phenomena or different aspects of the same phenomena. That is, allowing method factors to be correlated could contaminate them with correlations among indicators created by the substantial factors. As a consequence, this would lead to an overestimation of the relevance of method factors. Conversely, if fixing the correlations between method factors leads to an underestimation of the impact of method factors, then this means erring on the side of caution.

Confirmatory factor analysis does not only allow the estimation of factor loadings, variances, covariances, and correlations. More importantly, it allows the comparison of different models in terms of their fit to empirical data. Such model comparisons form the core of our research design. In order to assess the relevance of method factors, we conduct model comparison likelihood ratio tests with models which contain method factors against models which do not contain model factors. Such models can be obtained by deleting the term $\kappa_{ik}G_k$ from equa-

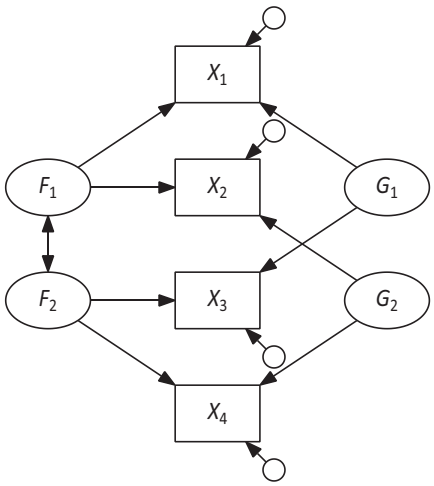

Figure 1. An Illustration of a factor model with conceptual factors and method factors.

[1] In factor analytic variants of multi-trait-multi-method analysis one would restrict the meaning of the term 'method factor' to common factors that represent the effects of a particular method of measurement. Since the influence of the creating institution on the regime indicators may be largely a consequence of the particular methods employed by this institution, we find it justifiable to use the term 'method factor' in a somewhat wider sense. If we had used a term like 'institutional factor' this might have led to confusion with conceptual factors that refer to institutional aspects of regime characteristics.

[2] Confirmatory factor analysis *per se* does not distinguish between different types of common factors. The distinction between two types of common factors, as reflected in different symbols used for the factors and their loadings, is a matter of interpretation that guides the construction of a factor model.

tion (1) and by deleting the nodes labelled G_1 and G_2 in Figure 1 as well as the arrows that connect them with the nodes labelled X_1, X_2, X_3, and X_4. The null hypothesis in these likelihood ratio tests is that a model without the method factors fits the data as well as a model that includes method factors. If a likelihood ratio test leads to the rejection of the null hypothesis, this means that an improvement of model fit brought about by the inclusion of method factors is more than a product of chance, which in turn provides evidence for the existence of method factors.

In order to make sure that our results are robust, we conduct our hypothesis tests based on the three different conceptualisations of the dimensionality of regime characteristics. That is, in the first variant the model that represents the null hypothesis has the two conceptual factors *Democratic rule* and *Political liberties*, in the second variant, the null model includes the three conceptual factors *Electoral principle*, *Liberal principle*, and *Participatory principle*, and in the third variant the null model includes the four conceptual factors *Civil rights*, *Electoral regime*, *Horizontal accountability*, and *Political rights*.

Apart from the identification problems that always lurk in complex CFA and structural equation models, our analysis is confronted with three related challenges: (1) non-normal and categorical indicators violate the standard assumptions on which likelihood-based inference in confirmatory factor analysis is based; (2) many indicators are conceptualised so that most democracies receive top scores—they are 'truncated'; (3) we observe the same countries at several points in time, which introduces dependencies not accounted for in standard CFA.

The first two challenges are illustrated by Figure 1: many indicators in our dataset place few countries in the centre of the empirical distribution and many at the extremes, in contrast to the shape of a normal distribution. Moreover, some of the indicators, in particular, the indicators from the Polity project, only have a small number of distinct values and therefore not have metric quality. In a situation like this maximum likelihood estimators may lose their asymptotic efficiency and test statistics may lead to false positives. For situations such as this Browne's (1984) asymptotically distribution-free estimator may retain asymptotic efficiency and provide relatively accurate test statistics. However, Browne's estimator requires a very large sample size, larger than the size of the data sets we employ in our analysis. For this reason, we stick to likelihood ratio test statistics and report Bollen-Stine bootstrap-based *p*-values (Bollen & Stine, 1992).

The second challenge is also illustrated by the histograms in Figure 2: contemporary democracies may be all too similar with respect to the regime indicators available in the data sets. Indicators such as Polity's *openness of executive recruitment* or V-Dem's *clean elections* show extreme peaks at the upper end of the scale. In order to address this problem, we repeat our analyses with subsets of the *D1* and *D2* data sets that contain only those regimes classified as non-democracies according to the democracy-and-dictatorship data (Cheibub et al., 2010; updated by Bjørnskov & Rode, 2017). We exclude all democracy regime types (codes 0 to 2) and retain all non-democracies (codes 3 to 5). Such a restriction of the data to non-democracies can eliminate some of the strongly peaked or U-shaped appearances in the histograms of the indicator variables, yet they still appear clearly non-normal.

The third challenge is that we use panel data, where measures are taken repeatedly from the same countries and therefore are not (conditionally) independent from one another. The methodology of CFA and structural equation modelling is mostly developed with cross-sectional data in mind, as is the available software to estimate such models. The (serial) dependence of measures taken from identical countries may or may not lead to biased estimates, but at least it will lead to inaccurate inference if standard errors are constructed based on the assumption of independence. We address this challenge with two approaches adopted from the econometrics of panel data (Baltagi, 2013): in a first approach, we fit our models to between country cross-sectional data constructed from the country-level means of the regime measures. This country-level aggregate data has a considerably smaller number of observations and thus a smaller power, but the serial dependence of the measures is eliminated. In a second approach, we keep the temporal information contained in the data and fit our models to within-country first differences of regime measures, i.e., to the amount of change compared to the previous year. This eliminates the between-country heterogeneity and reduces the serial dependence.

Considering four concepts with and without method factors means that we will have to estimate eight models for each of the two data sets we are employing (*D1* and *D2*), both for a full version of each data set and for the version reduced to non-democracies. In total, this gives 64 fitted models if we further distinguish between-country cross-sections and within-country first-differences. It is impossible to discuss the estimates based on this many model fits in a single article. Instead, we only discuss a series of chi-squared tests for model comparisons and present estimates only for models with four conceptual factors and four method factors fitted to data set *D2*. All models are estimated using the package *lavaan* (Rosseel, 2012) in the statistical environment *R* (R Core Team, 2017).

5. Results

As explained in the previous section, we conduct model comparison likelihood ratio tests to obtain evidence about the presence of method factors that represent the influence of the institutions on the regime indicators that they create. Each likelihood ratio test compares a model that contains only conceptual factors, common factors that represent only conceptual dimensions of regime

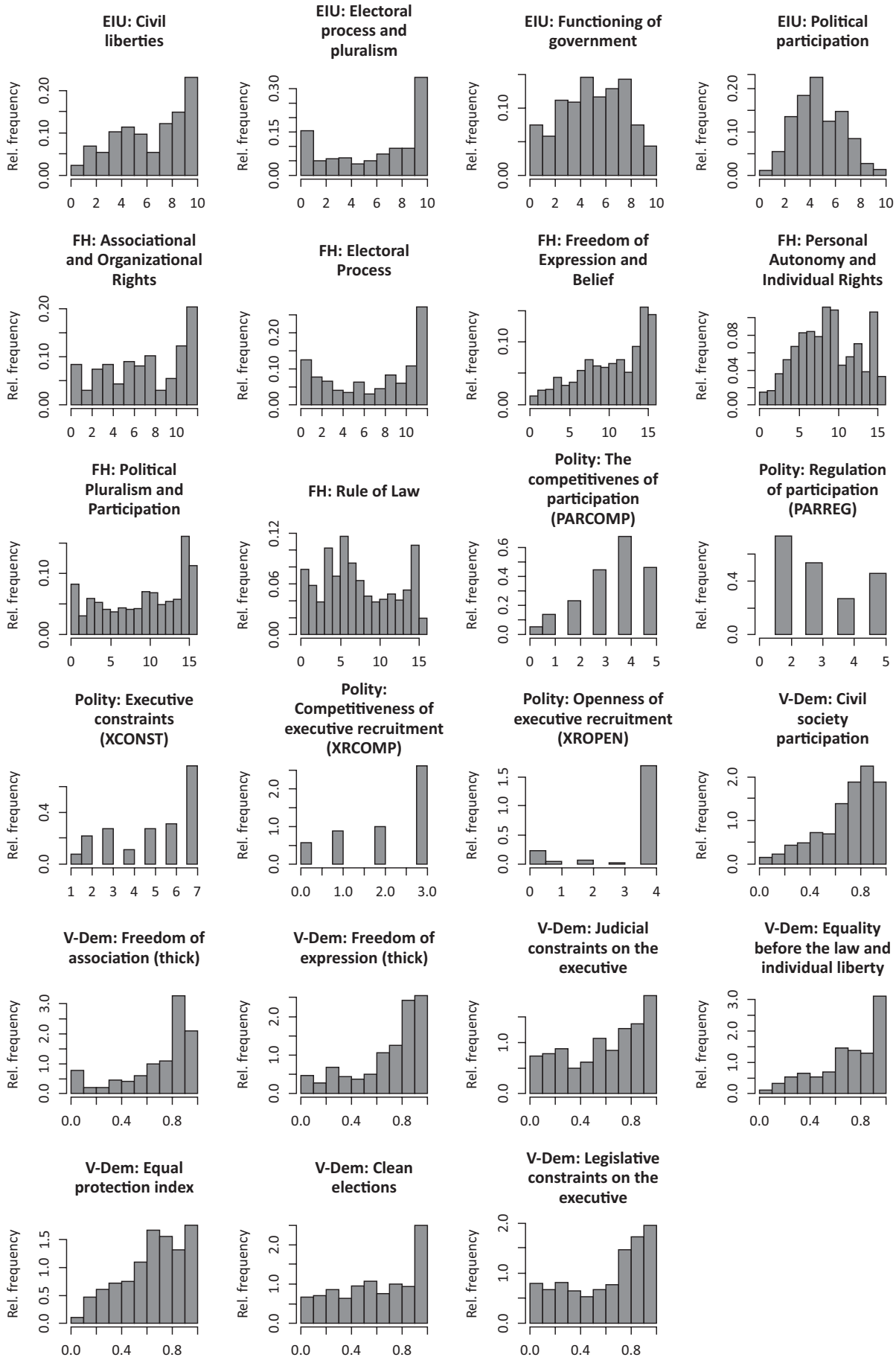

Figure 2. Histograms of the variables contained in *D2* (N = 1,070).

characteristics, with a model that additionally contains method factors corresponding to the various institutes that produced the indicators. If the likelihood ratio test indicates that inclusion of method factors leads to a statistically significant improvement of model fit, then we conclude that democracy measures indeed are affected by method factors. In order to make sure that our results are robust, we repeat the likelihood ratio test with different conceptualisations of regime dimensions as a baseline. Furthermore, we repeat the likelihood ratio tests with respect to the complete set of countries as well as only those countries categorized as authoritarian by Cheibub et al. (2010) and Bjørnskov and Rode (2017). To take into account the panel structure of the data, we conduct the tests first based on the between-country cross-section (i.e. the country averages) of the regime indicator values and second based on the within-country first differences of the indicator values. Table 2 shows the results of the hypothesis tests based on between-country cross-section while Table 3 shows the hypothesis tests results based on within-country first-differences. In addition to the values of the likelihood-ratio statistics, Tables 2 and 3 show three goodness-of-fit indices: the comparative fit index (CFI) which varies between 0 and 1, where 1 indicates perfect fit; the root mean square error of approximation (RMSEA), which also varies between 0 and 1, but where a value below 0,05 indicates an acceptable fit; and the standardised root mean squared residual (SRMR), which again usually varies between 0 and 1, with lower values indicating a better fit between the model and the data.

The results in Tables 2 and 3 are clear: no matter whether one assumes one, two, three, or four conceptual dimensions of regime properties, no matter whether one considers all countries or only non-democratic ones; the improvement of model fit by including method factors into the factor model is statistically significant at any conventional level. Furthermore, the goodness-of-fit indices also show a substantial improvement, except for SRMR in section (b) of Table 2. In summary, we find strong and robust evidence that method factors matter.

Having established the existence of method factors, we should discuss their relevance. How strong is their influence on measures of regime properties? This question can be answered by comparing the sizes of the loadings of the regime measures on the conceptual factors with their loadings on the method factors. In order to take into account the different scale lengths of the regime measures, such a comparison is best made using standardised estimates that are rescaled so that common factors and indicators all have unit variance. If the standardised loading of a regime measure on a method factor is as large as, or larger than, its loading on any conceptual factor then its validity should be considered questionable.

Figure 3 illustrates the factor loadings of the cross-section of regime indicators in the model that fit the data best, the factor model with four conceptual factors, and all four method factors. Diamonds represent conceptual factors, circles represent method factors, and rectangles represent regime measures. Each factor loading in the model is represented by an arrow, where the width indicates the absolute size of the standardised loading.[3] Overall, the

Table 2. Model comparison tests for the presence of method factors in data set *D2* (between-country cross-section).

(a) All countries

	Deviance	Mod.Df	Chi-squared	Diff. Df	p-value	CFI	RMSEA	SRMR
1 Dimension	1419.8	253				0.813	0.171	0.055
+ Method factors	891.9	230	527.9	23	0.000	0.894	0.135	0.045
2 Dimensions	1368.4	252				0.821	0.168	0.056
+ Method factors	844.0	229	524.5	23	0.000	0.901	0.131	0.045
3 Dimensions	1395.8	250				0.816	0.171	0.055
+ Method factors	859.4	227	536.5	23	0.000	0.898	0.133	0.045
4 Dimensions	1314.3	247				0.828	0.166	0.055
+ Method factors	822.0	224	492.3	23	0.000	0.904	0.130	0.045

(b) Non-democratic countries only

	Deviance	Mod.Df	Chi-squared	Diff. Df	p-value	CFI	RMSEA	SRMR
1 Dimension	695.1	253				0.697	0.177	0.102
+ Method factors	495.7	230	199.5	23	0.000	0.818	0.144	0.120
2 Dimensions	677.7	252				0.708	0.174	0.102
+ Method factors	460.6	229	217.1	23	0.000	0.841	0.134	0.114
3 Dimensions	693.7	250				0.696	0.178	0.101
+ Method factors	499.6	227	194.2	23	0.000	0.813	0.146	0.088
4 Dimensions	624.5	247				0.741	0.165	0.099
+ Method factors	438.9	224	185.6	23	0.000	0.853	0.131	0.104

[3] Path diagram of the confirmatory factor analysis model with four conceptual and for method factors—within-country first differences.

Table 3. Model comparison tests for the presence of method factors in data set *D2* (within-country first differences).

(a) All countries

	Deviance	Mod.Df	Chi-squared	Diff. Df	p-value	CFI	RMSEA	SRMR
1 Dimension	4256.9	253				0.503	0.132	0.105
+ Method factors	1478.5	230	2778.4	23	0.000	0.845	0.077	0.055
2 Dimensions	4236.5	252				0.505	0.132	0.106
+ Method factors	1478.2	229	2758.3	23	0.000	0.845	0.077	0.055
3 Dimensions	4151.3	250				0.515	0.131	0.104
+ Method factors	1403.1	227	2748.1	23	0.000	0.854	0.075	0.054
4 Dimensions	4111.2	247				0.520	0.131	0.104
+ Method factors	1340.0	224	2771.2	23	0.000	0.861	0.074	0.061

(b) Non-democratic countries only

	Deviance	Mod.Df	Chi-squared	Diff. Df	p-value	CFI	RMSEA	SRMR
1 Dimension	3123.6	253				0.424	0.113	0.100
+ Method factors	1140.7	230	1982.9	23	0.000	0.817	0.067	0.056
2 Dimensions	3106.7	252				0.427	0.113	0.102
+ Method factors	1140.5	229	1966.2	23	0.000	0.817	0.067	0.057
3 Dimensions	3065.7	250				0.435	0.112	0.100
+ Method factors	1126.4	227	1939.2	23	0.000	0.820	0.067	0.055
4 Dimensions	2977.3	247				0.452	0.111	0.102
+ Method factors	1016.8	224	1960.5	23	0.000	0.841	0.063	0.073

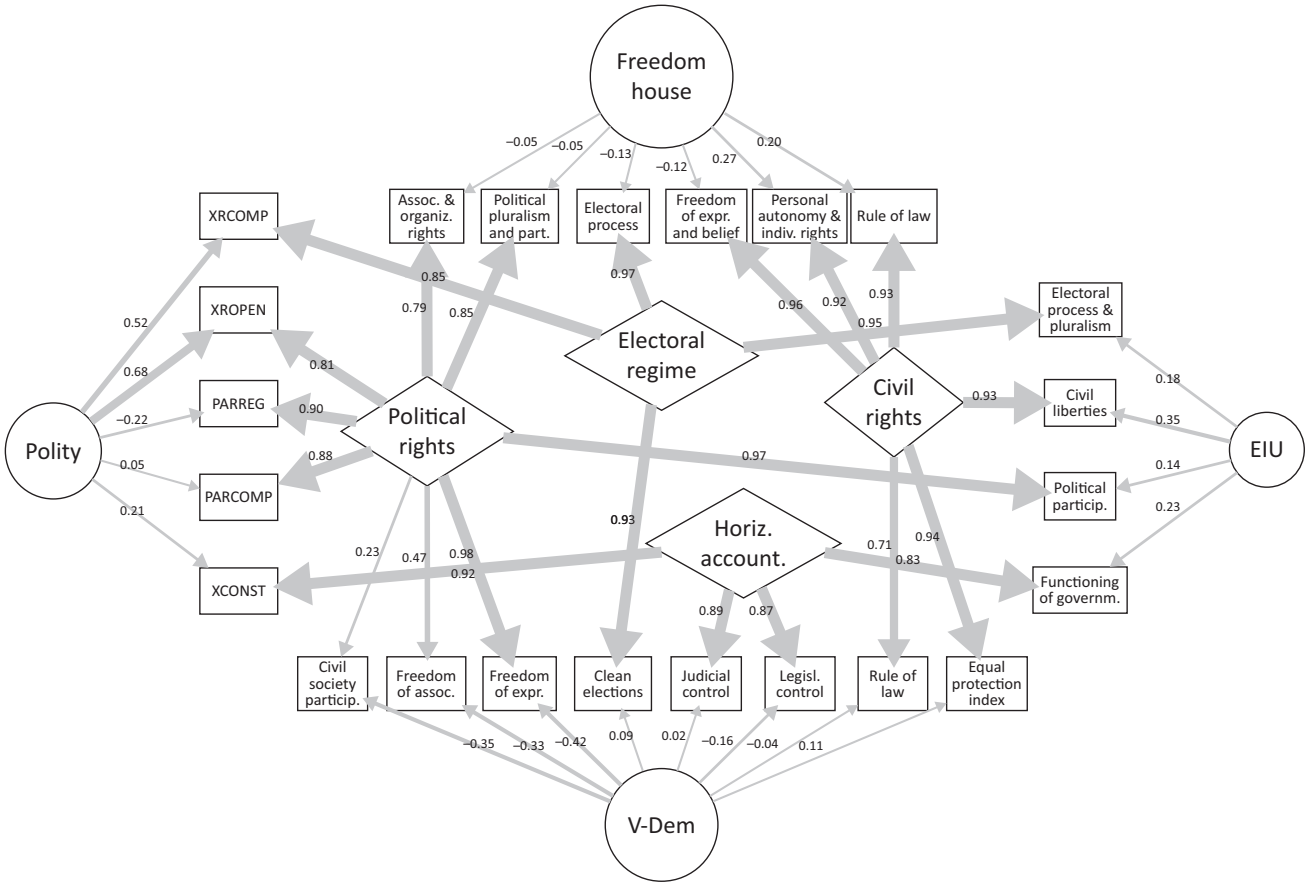

Figure 3. Path diagram of the confirmatory factor analysis model with four conceptual and for method factors—between country cross-section.

loadings of the regime measures on the conceptual factors are larger in absolute value than the loadings on the method factors. This is good news in so far as most regime measures indeed mostly represent substantial regime aspects. Yet, some of the loadings on the method factors are quite large. This affects, in particular, the regime measures from the Polity project: the indicators XRCOMP and XROPEN have strong loadings on the method factor. Even the relatively novel V-Dem measures do not seem to be without problems. All political rights indicators have relatively strong method factor loadings, and both *Civil society participation* and *Freedom of association* have small loadings on the conceptual factor. It appears that V-Dem contradicts more traditional sources on the relative position of countries in terms of political rights.

Figure 4, which illustrates loadings from a factor model fitted to within-country first differences, delivers an even less comforting message: the loadings on the method factors appear at least as large as the loadings on conceptual factors, thus raising doubts regarding the validity of many of the regime measures—at least when it comes to adequately representing change of regime properties within a country.[4] In particular, the Freedom House measures appear to be much more affected by the corresponding method factor than by any of the conceptual factors. Also, some of the Polity measures show very strong loadings on the method factor. In general, these strong loadings indicate that there is much less consensus between the various sources in terms of change than in terms of the average character of a country. This appears to affect the *Political rights* and *Civil rights* factors in particular, and much less so the *Electoral regime* and *Horizontal accountability* measures. A potential explanation for this pattern is that the latter refer to institutional characteristics that are rather easily observable in the form of laws and regulations, while changes of (effective) civil and political rights are unobservable latent properties and therefore more prone to follow a source's bias.

How can we make sense of the divergent assessments of method factors in democracy indicators that the two Figures suggest? One interpretation is that the different producers of democracy indicators vary in their sensitivity to change within countries—some adjust indicators earlier, others later (cf. Lueders & Lust, 2017). Method factors are less salient when it comes to country averages because the producers eventually converge to similar assessments once the dust raised by changing regime properties has settled. An alternative interpreta-

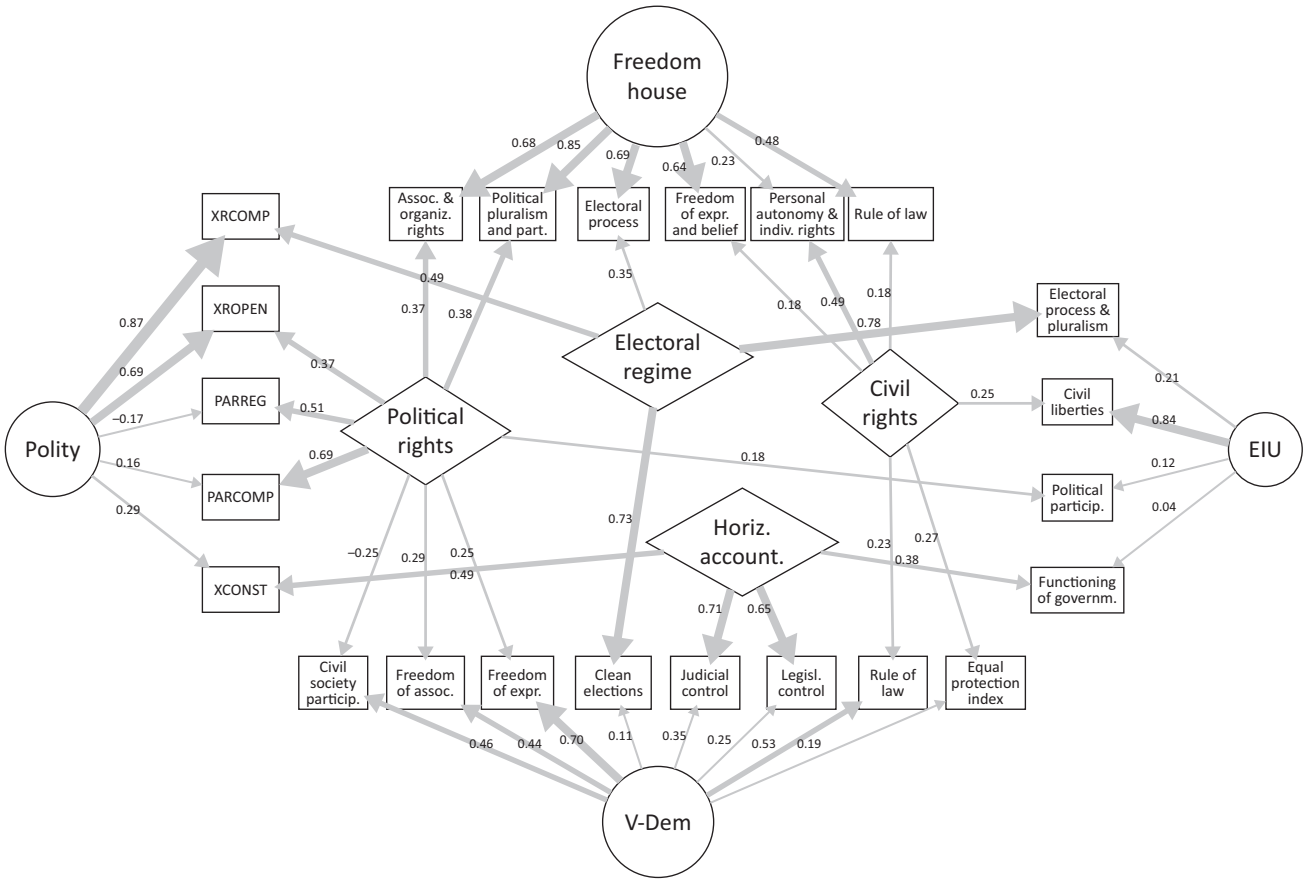

Figure 4. Path diagram of the confirmatory factor analysis model with four conceptual and for method factors—within-country first differences.

[4] One should not be misled by the all negative loadings on the Civil rights factor. This is empirically equivalent with a model fit where all loadings on this factor (and also the covariances with the other factors) have their signs reversed.

tion of the small loadings on the dimension factors in the within-country models is that within-country changes (in particular in civil and political rights) occur not all at the same time but in waves. As a consequence, the covariances between first differences would understate the actual patterns of changes. Alas, the large sizes of method factors make such a benevolent interpretation less likely.

Moreover, the temporal dependence of indicators from the same source has a plausible explanation: if a data producer comes to the conclusion that a larger regime change is occurring, they may also form the expectation that several indicators portraying different aspect of the regime would change at once. A desire to maintain consistency in regime measures may thus increase the correlation between indicators created by the same producer and decrease the correlation with regime measures created by other data producers.

6. Conclusion and Recommendations

Having analysed democracy indicators from four different sources, we find strong evidence for systematic deviations, i.e., method factors. The question of whether these method factors constitute biases in all cases, or whether one source is simply closer to measuring 'true' democracy cannot be answered here. We can state that while most sources converge on cross-sectional variation, they diverge on temporal variation within countries. Much of this uncertainty is not random error affecting individual indicators, but systematic error driven by the source.

As we analyse cross-sectional and within-country variation separately, we can render our speculations on the origins of this systematic error more precisely. Sources agree much more on the cross-sectional data than on the within-country differences.[5] This is a picture that could be explained by a practice of 'guessing until convergence'. Measuring change in democracy is difficult—in particular in the 'softer' dimensions of political and civil rights. When one data producer, at a certain point in time, perceives a change in a country, but others do not, agreement on the affected dimension declines. If those perceived changed stand the test of time, other data producers will follow and also adjust their scores. If the democratic practice observed in the country does not seem to warrant the change in scores, the first mover will revert their scores. As a result, we see substantial agreement in average cross-sectional assessments over the entire time period that we investigate. But we do not see much agreement on the timing of changes. It would be interesting to investigate whether a particular source is better at predicting change—a potential 'superforecaster' among democracy index producers.

Looking at individual sources, the largest method factors can be observed for some of the Polity IV indicators. This may in part be explained by the divergent conceptual setup of the Polity indicators: in the Appendix, we discuss the assignment of the indicators to the conceptual dimensions, and these decisions are more ambiguous for the Polity indicators. For example, Polity focuses more on the logic of ruler selection and less on participation, as exemplified by the neglect of suffrage issues (Munck & Verkuilen, 2002, p. 11). Freedom House exhibits less bias on the cross-section, but it on average it fares worst of all sources when assessing changes. In this light, Freedom House's self-declared mission 'to defend human rights and promote democratic change',[6] could inspire speculation that temporal distortions in Freedom House data are indeed intentional, with the aim to spur regime change. Previous studies have confirmed an ideological bias of this source (Giannone, 2010; Steiner, 2016). There is less critical literature on V-Dem yet, as most published work using the dataset comes from the large project team itself. The V-Dem team has also shown a large effort to assess and improve the quality of their data. For example, it has presented a comparison of its aggregate *polyarchy* score (a summary measure of electoral democracy) with other high-level indices (Teorell, Coppedge, Skaaning, & Lindberg, 2016, pp. 28–31). Nonetheless, some V-Dem indicators exhibit sizable method factors and should be investigated. We can say little about the EIU's democracy index beyond a diagnosis of moderate method factors, as hardly any complementary research on this source is available.

For applied researchers, our results shall serve as a reminder to adhere to some well-known but not always heeded rules of good practice. In a nutshell, these are (1) use the best source available, (2) use several sources, and (3) use meta indices. Determining the best sources available always depends on the research question at hand. An indicator with high conceptual validity for a particular application should certainly not be replaced with a more reliable measure that is far less valid. Contextual specificity matters (Adcock & Collier, 2001, p. 534). Among our set of indicators, however, we have many close matches that claim to measure very similar issues. In that case, given our results, our best guess for the indicator least affected by method bias will usually be the indicator provided by the V-Dem project. This is based not only on our model estimates but also on what we know about how the data is generated.

This scenario leads us to the second recommendation: should several indicators be available, use them! There is little additional effort in using multiple indicators to assess the robustness of results. Data collections such as those published by the *Quality of Government Institute* provide a large variety of indicators merged ready for the end user.

The third recommendation requires more preparation: the use of meta indices. For democracy as a unidimensional concept, various estimates exist. A prominent example based on a Bayesian measurement model is the

[5] Also note that the failure to agree on changes is all the more disappointing since we are employing yearly data, not monthly or weekly assessments.
[6] Available at https://freedomhouse.org/our-work

Unified Democracy Scores (Pemstein et al., 2010). For sub-dimensions of democracy, there are fewer meta indices available. Examples beyond Bollen's (1990) original approach are the contestation and participation scores provided by Coppedge, Alvarez and Maldonado (2008). In order to provide more choice on meta indices for sub-dimensions of democracy, one could employ the very factor scores that our models produce. This would provide quantitative measures for the dimensions of the Bollen concept, the V-Dem concept, and the Merkel concept. However, before using these in applied research, comprehensive additional vetting will be required, as validly measuring a substantial concept is more demanding than validly detecting method bias.

Our advice to producers of democracy indicators who pursue the goal of unbiased measures of democracy is to further address issues of method bias along all stages of the measurement process with various methodological approaches (see McMann, Pemstein, Seim, Teorell, & Lindberg, 2016 for an example) and with reference to alternative sources. Coppedge et al. (2017a), for example, have taken first steps to compare V-Dem to its main competitors. Additional efforts to more precisely assess temporal change in unobservable traits of democracy are advised.

Acknowledgments

We thank the editors of this thematic issue, two anonymous referees, participants at the Annual Meeting of the Methods Section of the German Political Science Association in May 2017, and participants at the V-Dem Lunch Seminar in September 2017 for comments and suggestions. We acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme Open Access Publishing by the Baden-Württemberg Ministry of Science, Research and the Arts and by Ruprecht-Karls-Universität Heidelberg.

Conflict of Interests

The authors declare no conflict of interests.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529–546.

Baltagi, B. H. (2013). *Econometric analysis of panel data* (5th ed.). Chichester: Wiley.

Beck, T., Clarke, G., Groff, A., Keefer, P., & Walsh, P. (2001). New tools in comparative political economy: The database of political institutions. *The World Bank Economic Review*, *15*(1), 165–176.

Bertelsmann Transformation Index. (2016). *Transformation index BTI 2016: Political management in international comparison*. Gütersloh: Bertelsmann Stiftung.

Bjørnskov, C., & Rode, M. (2017). *Regime types and regime change: A new dataset*. Aarhus: Aarhus University.

Bollen, K. A. (1990). Political democracy: Conceptual and measurement traps. *Studies in Comparative International Development*, *25*(1), 7–24.

Bollen, K. A. (1993). Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science*, *37*(4), 1207–1230.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

Cheibub, J. A., Gandhi, J., & Vreeland, J. R. (2010). Democracy and dictatorship revisited. *Public Choice*, *143*(1/2), 67–101.

Coppedge, M., Alvarez, A., & Maldonado, C. (2008). Two persistent dimensions of democracy: Contestation and inclusiveness. *The Journal of Politics*, *70*(3), 632–647.

Coppedge, M., Gerring, J., Lindberg, S. I., Skaaning, S.-E., & Teorell, J. (2017a). *V-Dem comparisons and contrasts with other measurement projects* (Varieties of Democracy Working Paper Series, 2017(45)). Gothenburg: V-Dem Institute.

Coppedge, M., Gerring, J., Lindberg, S. I., Skaaning, S.-E., Teorell, J., Altman, D., . . . Wilson, S. (2017b). *V-Dem country-year dataset v7*. Gothenburg: V-Dem Institute.

Dahl, R. A. (1971). *Polyarchy: Participation and opposition*. New Haven, CT: Yale University Press.

Economist Intelligence Unit. (2014). *Democracy index 2014: Democracy and its discontents*. Retrieved from http://graphics.eiu.com/PDF/Democracy_Index_2010_web.pdf

Freedom House. (2016). *Freedom in the world 2016*. Retrieved from https://freedomhouse.org/report/freedom-world/freedom-world-2016

Giannone, D. (2010). Political and ideological aspects in the measurement of democracy: The freedom house case. *Democratization*, *17*(1), 68–97.

Giebler, H. (2012). Bringing methodology (back) in: Some remarks on contemporary democracy measurements. *European Political Science*, *11*(4), 509–518.

Lijphart, A. (2012). *Patterns of democracy: Government forms and performance in thirty-six countries* (2nd ed.). New Haven, CT: Yale University Press.

Lindberg, S. I., Coppedge, M., Gerring, J., & Teorell, J. (2014). V-Dem: A new way to measure democracy. *Journal of Democracy*, *25*(3), 159–169.

Lueders, H., & Lust, E. (2017). *Multiple measurements, elusive agreement, and unstable outcomes in the study of regime change* (Varieties of Democ-

racy Working Paper Series, 2017(52)). Gothenburg: V-Dem Institute.

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2016). *Polity IV project: Political regime characteristics and transitions, 1800–2015: Dataset users' manual*. Vienna, VA: Center for Systemic Peace. Retrieved from http://www.systemicpeace.org/inscr/p4manualv2015.pdf

McMann, K. M., Pemstein, D., Seim, B., Teorell, J., & Lindberg, S. I. (2016). *Strategies of validation: Assessing the Varieties of Democracy corruption data* (Varieties of Democracy Working Paper Series, 2016(23)). Gothenburg: V-Dem Institute.

Merkel, W. (2004). Embedded and defective democracies. *Democratization*, *11*(5), 33–58.

Munck, G. L., & Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative Political Studies*, *35*(1), 5–34.

O'Donnell, G. A. (1994). Delegative democracy. *Journal of Democracy*, *5*(1), 55–69.

Pemstein, D., Meserve, S. A., & Melton, J. (2010). Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis*, *18*(4), 426–449.

R Core Team. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Retrieved from https://www.R-project.org

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

Steiner, N. D. (2016). Comparing Freedom House democracy scores to alternative indices and testing for political bias: Are US allies rated as more democratic by Freedom House? *Journal of Comparative Policy Analysis: Research and Practice*, *18*(4), 329–349.

Teorell, J., Coppedge, M., Skaaning, S.-E., & Lindberg, S. I. (2016). *Measuring electoral democracy with V-Dem data: Introducing a new polyarchy index* (Varieties of Democracy Working Paper Series, 2016(25)). Gothenburg: V-Dem Institute.

Teorell, J., Dahlberg, S., Holmberg, S., Rothstein, B., Khomenko, A., & Svensson, R. (2017). *The quality of government standard dataset, version jan17*. University of Gothenburg: The Quality of Government Institute. Retrieved from http://www.qog.pol.gu.se doi:10.18157/QoGStdJan17

Treier, S., & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*(1), 201–217.

About the Authors

Martin Elff is a professor of political sociology at Zeppelin University, Friedrichshafen. He holds a doctorate in social science (Dr. rer. soc.) from the University of Mannheim. His research interests are in political attitudes and political behaviour, party competition, and quantitative methods of political science.

Sebastian Ziaja is a postdoctoral researcher at the Research Center for Distributional Conflict and Globalization at Heidelberg University. He holds a PhD from the Department of Government at the University of Essex. His research focusses on political regimes, democracy aid, state fragility, and the measurement of social science concepts.