European Political Science (2021) 20:666–684
https://doi.org/10.1057/s41304-021-00328-8

DATASET

Comparing measures of democracy: statistical properties, convergence, and interchangeability

Andrea Vaccaro[1]

Accepted: 11 February 2021 / Published online: 16 March 2021
© The Author(s) 2021

Abstract

By comparing and analysing four cross-national measures of democracy, this article provides novel information regarding the statistical properties, convergence, and interchangeability of some of the most frequently used measures of democracy. The author points out limitations related to the statistical properties of these measures and finds that even if measures of democracy are highly convergent, their interchangeability is weak. This means that the choice of the measure of democracy has considerable consequences for the conclusions of a given study. Especially so in studies covering the last few decades, because the author finds that in general the interchangeability of democracy measures has decreased since the 1980s. In choosing one measure over another, scholars should be aware of the limitations identified in this article. To overcome problems related to weak interchangeability, if a single measure cannot be credibly chosen on theoretical grounds, the author recommends users of the measures to validate their findings with multiple measures of democracy.

Keywords Correlation · Democracy · Democratisation · Indicators · Interchangeability · Measurement

Introduction

Democracy, its causes, and its consequences are among the most extensively analysed issues in political research. Nevertheless, many theoretical propositions on democracy remain contested because of inconsistency in empirical findings. This paper argues that one underemphasised reason to inconclusive findings on the topic can be attributed to the use of different democracy measures and provides strong empirical support for the argument. Differences in measures of democracy

✉ Andrea Vaccaro
andrea.vaccaro@uniroma1.it

[1] Department of Social Sciences and Economics, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy

https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

are problematic, because scholars have more and more measures of democracy to choose from, but choosing the right measure and/or assessing whether the findings are generalisable across several measures does not play an important role in most research agendas.

The limitations of measures of democracy have been discussed in a rich body of literature, and scholars have identified problems related to reliability and validity (e.g. Bollen 1980; Bowman et al. 2005; Hadenius and Teorell 2005; Knutsen 2010; Giebler 2012; Pickel et al. 2015). More specifically, to give a few examples, some of the most established measures of democracy lack in conceptual clarity and contextual equivalence (Barsh 1993), are volatile over time (Gleditsch and Ward 1997), have systematic measurement errors (e.g. Bollen and Paxton 2000; Treier and Jackman 2008; Pemstein et al. 2010; Elff and Ziaja 2018), little theoretical background (Besançon 2003), and ideological bias (Giannone 2010). Elkins (2000) finds that dichotomous measures of democracy are less valid and reliable than graded measures. Munck and Verkuilen (2002) distinguish among three major methodological problems of democracy indices: conceptualisation, measurement, and aggregation. McMahon and Kornheiser (2010) show that not all democracy indices are strongly related to each other. Boese (2019) argues that the V-Dem performs better than Freedom House and Polity IV as to definition, measurement scale, and aggregation.

Casper and Tufis (2003) and Högström (2013) investigate whether measures of democracy are interchangeable by regressing selected measures of democracy on structural predictors of democracy. Through panel (Casper and Tufis 2003) and cross-sectional (Högström 2013) regressions, both studies find that measures of democracy are not interchangeable and that the selection of measures can determine significantly the outcomes of a study. Both studies, however, have some limits that are addressed in my research. First and foremost, Casper and Tufis (2003) evaluate the interchangeability of three democracy indices (Vanhanen, Polity IV, and Freedom Rating) from 1975 to 1992 and Högström (2013) the interchangeability of two democracy indices (Polity IV and Freedom Rating) in 2009. It is evident that a more comprehensive picture on the matter requires analysing a larger number of democracy measures in multiple years. Moreover, updated information on more recent years is needed. My research builds primarily on these two studies with three key improvements: (1) a comparative evaluation of the statistical properties of the measures, (2) an analysis of the overall and over time convergence of the measures, and (3) broader empirical foundations covering more recent data and some of the most "currently relevant" measures of democracy.

Data and research strategy

Selection of relevant measures of democracy

Since all existing measures of democracy cannot be effectively compared and analysed in a single study, I begin my enquiry by selecting some of the most currently relevant measures of democracy. The selection is based on academic importance, measurement scale, availability, and coverage. Dichotomous indicators such as



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Alvarez et al. (1996) *Democracy and Dictatorship* are excluded, so that we capture not only the distinction between autocracy and democracy, but also different degrees of democracy. Ordinal measures that are hardly ever treated as interval such as Skaaning et al. (2015) *Lexical Index of Electoral Democracy* are excluded as well. Finally, measures that have limited time coverage such as Economist Intelligence Unit's (2019) *Democracy Index* have not been updated regularly in recent years such as Coppedge et al. (2008) *Contestation* and *Inclusiveness*, or rely explicitly on other measures such as Kaufmann et al. (2011) *Voice and Accountability* are left out from the analysis.

To assess the academic importance of measures of democracy, I use Coppedge et al. (2017) study, which reports the number of Google Scholar citations of several democracy indices. Center of Systemic Peace's *Polity IV* leads the pack with nearly 5000 citations. Freedom House's *Civil Liberties* and *Political Rights* are the second most "impactful" measures of democracy with more than 1,500 citations on Google Scholar. Vanhanen's *Competition* and *Participation* are the next most cited measures of democracy that fulfil our selection criteria. The three above-mentioned measures have been widely used in political research, cover nearly all countries around the world over many decades, and are publicly available. Besides these three well-established measures, I include also the more recently developed V-Dem's *Polyarchy* index, which has gained widespread recognition the last few years. Thus, to sum up, this article analyses four measures of democracy: Polity IV (Marshall et al. 2019), Vanhanen Index (Vanhanen 2019), Freedom Rating (Freedom House 2019a), and V-Dem Polyarchy (Teorell et al. 2019). Table 1 presents an overview of these indices.

Polity IV (PIV) was originally built upon Gurr's (1974) and Eckstein and Gurr's (1975) studies on political systems, with the objective of analysing whether common beliefs about the characteristics of state authorities have possible historical or cross-cultural patterns around the world. The Polity IV index is composed of the Institutional Democracy and Institutional Autocracy sub-indices, which measure, respectively, the democratic and the autocratic features of a country. The two sub-indices, which in turn are based on several more specific sub-components, are ultimately combined into the Polity IV index. The index has a 21-point scale from more autocratic to more democratic and provides annual data for all countries in the world with a population of at least 500,000 since 1800.

Vanhanen's index of democratisation (VAN) measures democracy as a two-dimensional construct formed by the amount of electoral competition between parties and the amount of voting turnout in the elections (Vanhanen 2000). The index aggregates two equally weighted sub-indicators, Electoral Competition and Political Participation, which are both based on "hard" quantitative electoral data. Thus, according to Vanhanen (2000) the index provides more objective information than many other democracy indices. In theory, Vanhanen's index ranges from 0 to 100 (from low to high democracy), although, in practice, no country has ever received a score higher than 49. Annual data is provided for most countries in the world since 1810.

Freedom Rating (FH), developed and produced initially by regional studies specialist Gastil, is conceptually rooted in the principles of the Universal Declaration



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Table 1 Main characteristics of frequently used measures of democracy

Measure	Years	Countries	Scale	Underlying components
Vanhanen Index of Democracy (VAN)	1810–2018	180	0–100	Political competition and political participation
Freedom House Rating (FH)	1972–2019	209	1–7	Electoral process, political pluralism and participation, functioning of government, freedom of expression and belief, associational and organisational rights, rule of law, and personal autonomy and individual rights
V-Dem Polyarchy (VDEM)	1789–2019	179	0–1	Freedom of association, clean elections, freedom of expression, elected officials, and suffrage
Polity IV (PIV)	1800–2018	167	− 10 to 10	Political participation, executive recruitment, and constraints on chief executive

Note: Number of countries refers to year 2017



of Human Rights of 1948 (Freedom House 2019b). The index aggregates two main sub-indices, Political Rights and Civil Liberties, which in turn are based on several more specific sub-indicators. The final Freedom Rating score of a country is the average of these two main sub-indices. Annual data is available since 1972 for virtually all existing countries and some territories (e.g. Hong Kong, West Bank) in the world. The index measures democracy on a scale from 1 to 7, where a lower score stands for a higher level of democracy.

V-Dem's Polyarchy index (VDEM), rooted in Dahl's (1971) influential theory of democracy, quantifies democracy as a combination of "freedom of association, suffrage, clean elections, elected executive, and freedom of expression" (Coppedge et al. 2017: 10). In V-Dem project's conceptual scheme, the Polyarchy index is conceived to measure the fundamental aspects of democracy (Coppedge et al. 2017). The index synthesises 40 sub-indicators that are coded by country experts (mainly academics) and research assistants (Teorell et al. 2019). Polyarchy ranges on a continuous scale from 0 to 1, from low to high. Annual data covers nearly all countries of the world since 1789.

Research strategy

I start the quantitative analysis by exploring the statistical features and evaluating the distributional characteristics of the selected indices through violin plots. While we should not expect measures of democracy to follow a Gaussian distribution by nature, we would prefer—for statistical purposes—to have measures that are more or less normally distributed, for the simple reason that many common statistical analyses and tests assume such a distribution. An analysis of the distributions is an informative task per se. Yet, it evidences also some more subtle weaknesses in the measures that are further discussed. In particular, the ability to differentiate between countries (Knutsen 2010), especially at the extremes of a scale (Goertz 2020), is a desirable property of democracy measures. Violin plots are especially helpful in detecting bumps within a distribution (Hintze and Nelson 1998), and as we shall see, some of the democracy measures tend to clump at extreme values. Such scales should be further extended to capture all interesting variation (Goertz 2020).

Second, after analysing the statistical features and distributions of the data, I run bivariate correlations to assess the strength of the linear associations among our measures of democracy. Since all measures quantify democracy, we expect them to be strongly related to one another. To get a general view of the associations, correlations are computed between pairs of democracy measures using all available observations since 1972.[1] To examine whether these relationships differ from year to year and/or over time, correlations are run also at 5-year intervals. These correlations provide information about the statistical association and the convergence of the measures. If currently relevant measures of democracy are strongly associated among one another, we can conclude that their convergence is high.

[1] This year coincides with the first available year of FH and thus is selected as the first year of analysis.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Finally, through a battery of regression models, I analyse whether the choice of the measure of democracy affects the interpretation of the results, or in other words, whether measures of democracy are interchangeable or not. Following previous empirical literature on measurement interchangeability (e.g. Casper and Tufis 2003; Møller and Skaaning 2011; Högström 2013), I regress our measures one by one on the same set of predictors, keeping everything else equal. According to the interchangeability criterion, "equivalent measures should produce similar causal inferences" (Seawright and Collier 2014: 124). Hence, if the conclusions are not determined by the choice of the measure, we can claim that measures of democracy are interchangeable. To exclude selection bias, I run the regressions with only country-years common to our four measures of democracy.

The democracy indices are regressed on a set of common predictors of democracy based on both Högström's (2013) and Casper and Tufis' (2003) studies. These common or "structural" predictors are economic development, growth, trade openness, inflation, population, education, presidentialism, parliamentarianism, and party fractionalisation. The first five variables are taken from the World Development Indicators dataset (World Bank 2017). Economic development is measured by the natural logarithm of GDP/capita in constant 2010 USD. Growth is measured by the annual GDP/capita growth (as a percentage). Trade openness is measured by the sum of exports and imports of goods and services as a share of GDP. Inflation is measured by annual change in consumer prices (as a percentage). Population is measured by the natural logarithm of the estimated total population. Education is measured by the average years of education among citizens older than 15 and is retrieved from the Varieties of Democracy dataset (Coppedge et al. 2019). Party fractionalisation is measured by the probability that two random draws from the lower legislative chamber will be from the same party and is taken from Henisz's (2017) POLCON dataset. Presidentialism and parliamentarism are dummy variables coded according to Anckar and Fredriksson's (2019) classification of political systems.

The selected set of variables allows us to assess the interchangeability of measures of democracy with a sample of nearly 130 countries from 1973 to 2017. To cope adequately with problems common to panel regressions in cross-national contexts, as recommended by Beck and Katz (1995), I estimate the models with ordinary least squares (OLS) and panel-corrected standard errors (PCSE), which correct for heteroskedasticity and contemporaneous correlation. Serial correlation is corrected with a lagged dependent variable on the right-hand side of the equation and problems of reverse causality are mitigated by lagging all independent variables by 1 year. Throughout the empirical analysis, to facilitate comparison, measures of democracy are normalised (min–max) and transformed to range from 0 (low) to 100 (high). FH is reversed so that higher scores indicate more democracy.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

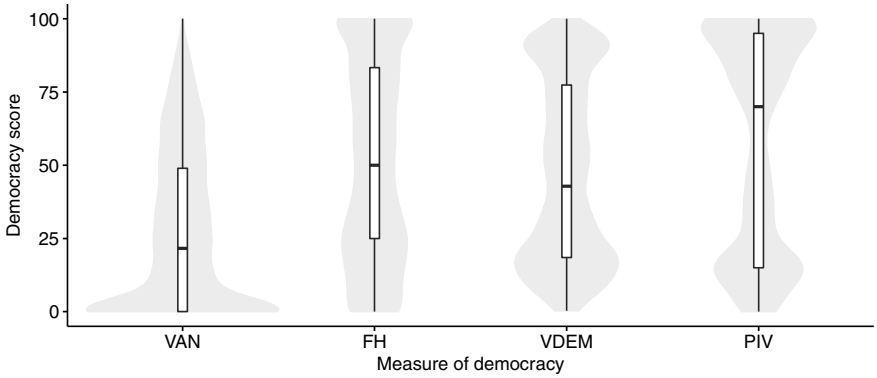


Fig. 1 Violin plots of measures of democracy (1972–2018)

Results and discussion

Main statistical properties

Violin plots in Fig. 1 present the main statistical properties of our four measures of democracy in all common country-years from 1972 to 2018. The black rectangle in the middle of each "violin" represents the median, the white box around it represents the interquartile range, and the whiskers show the distribution of the remaining values, with the exception of possible outliers. The grey area around the white box illustrates the shape of the distribution at various levels of democracy.

Overall, none of the selected measures seems to provide anything close to a fairly normal distribution of values. VAN has a substantially lower median value compared to the other measures and its distribution is heavily skewed to the right. In practice, as shown by the high frequency of observations at the bottom of the scale, this means that a large amount of country-years in VAN are coded as highly autocratic. On the contrary, the other three indices of democracy have a more or less pronounced bimodal distribution, with peaks near the two ends of the scale and less observations in the middle of the scale. Intermediate levels are particularly empty with PIV, as shown by its comparatively large interquartile range. The very extremes are more clumped with FH and PIV than with VDEM. FH's median is precisely in the middle of the scale, PIV's median is closer to the democratic end of the scale, and VDEM's median is closer to the autocratic end of the scale.

Besides showing that measures of democracy are far from being Gaussian, the distributional characteristics reveal some interesting differences among the measures. VAN seems to be unable to capture variation at the bottom extreme of the scale, as confirmed by the fact that as many as 25.7% of all country-years from 1972 to 2018 have the minimum score with VAN. Only in 2000, there are nearly 30 countries with the minimum possible level of democracy with VAN. Since the other measures are able to tell the difference between almost all these countries (VDEM) or at least between many of them (PIV and FH), such an agglomeration of country-years at the bottom extreme of the scale with VAN is more likely to reflect an



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Table 2 Pairwise correlation coefficients of measures of democracy (1972–2018)

	VAN	FH	VDEM	PIV
Vanhanen index (VAN)	1	0.84 (7421)	0.90 (7436)	0.87 (7181)
Freedom Rating (FH)	0.83 (7421)	1	0.90 (7343)	0.89 (7088)
Polyarchy (VDEM)	0.87 (7436)	0.92 (7343)	1	0.92 (7214)
Polity IV (PIV)	0.82 (7181)	0.88 (7088)	0.91 (7214)	1

Note: Upper-right quadrant: Spearman's correlations; bottom-left quadrant: Pearson's correlations; number of observations in parentheses; all coefficients significant at the $p < 0.001$ level

excessive compression rather than real equivalence between the observations. Furthermore, since all these countries have the lowest possible score in VAN, the index is not able to capture any decrease in the level of democracy in suspiciously many countries.

To a lesser extent, PIV and FH suffer also from similar distributional shortcomings at the very extremes of the scale. Both indices compress several countries especially at the maximum level of democracy, suggesting that the scales of PIV and FH should be further extended beyond their current upper extremes to capture a more nuanced view of highly democratic countries. In total, from 1972 to 2018, 18.6% of country-years in PIV and 14.1% country-years in FH are coded with the maximum score, and for instance, both indices are unable to distinguish between more than 30 countries with maximum scores each year from 2010 to 2015. By construction, any possible increase in democracy in these countries has been left unnoticed by both PIV and FH. Furthermore, as before, we can confidently assume that such an agglomeration of observations reflects excessive compression instead of true equivalence, because both VAN and VDEM are able to distinguish between the vast majority of these countries.

The above discussed violin plots show that none of the measures of democracy is normally distributed. Heavy non-normality is not problematic in itself but becomes limiting if users of these measures wish to employ them in statistical analyses and tests that assume normally distributed data. The evaluation of the distributional characteristics of the data shows as well that some of the measures fail to capture variation between non-democratic cases at the bottom end of the scale (VAN) or between highly democratic cases at the top end of the scale (PIV and FH). Such an artificial agglomeration of cases at the extremes of the scale can be considered as a shortcoming of VAN, FH, and PIV. From this perspective, VDEM is better constructed than the other measures.

Convergence

Table 2 reports bivariate correlations among our four democracy indices, when all available observations across countries and over years are taken into account. The bottom-left quadrant reports Pearson's correlation coefficients, while the upper-right quadrant reports Spearman's rank correlation coefficients. Both methods are used to add robustness to the results. Pearson's correlations measure the strength of linear



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

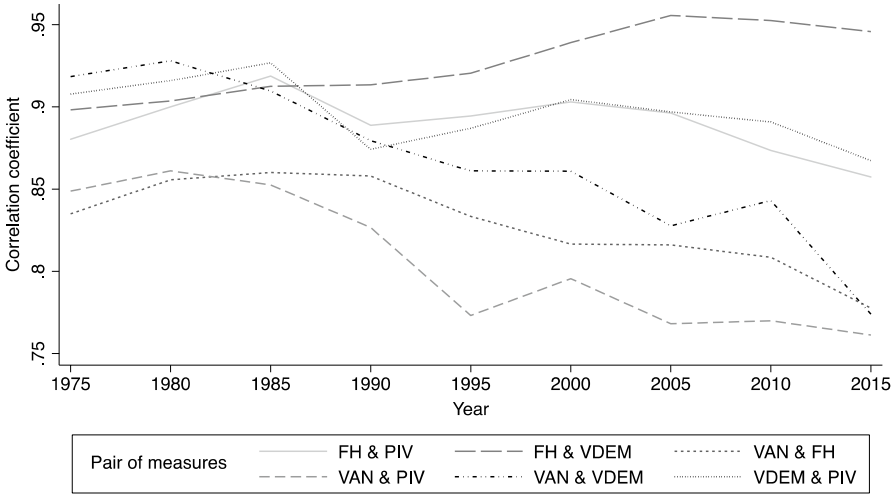


Fig. 2 Evolution of correlations over time: line plot of linear bivariate correlations among measures of democracy in 5-year intervals, 1975–2015

relationships and are more sensitive to outliers, while Spearman's correlations measure the strength of monotonic relationships and are less sensitive to outliers. The choice of the estimator, however, does not alter substantially the results.

All bivariate correlation coefficients, regardless of the chosen method, are higher than 0.80, indicating high convergence among measures of democracy. With Pearson's method the weakest correlation is between VAN and PIV (0.82), while the strongest is between VDEM and FH (0.92). With Spearman's method, instead, the weakest correlation is between VAN and FH (0.84) and the strongest is between VDEM and PIV (0.92).

In general, the high convergence of measures of democracy is confirmed by Fig. 2, which illustrates bivariate Pearson's correlation coefficients among our indices at 5-year intervals, from 1975 to 2015. Despite being strong, however, these correlations are not always constant over time. For instance, the bivariate correlation between FH and PIV changes minimally from one period to another (0.88 in 1975 and 0.86 in 2015), whereas the bivariate correlation between VAN and VDEM changes considerably over time (0.92 in 1975 but only 0.77 in 2015). In most cases, measures of democracy have become less similar among one another, as shown by the predominantly decreasing lines in the plot. Heterogeneity among correlation coefficients has increased over time. In 1975, the correlation coefficients among measures of democracy range only between 0.84 and 0.92, but in 2015, they range no less than from 0.76 to 0.95.

The findings of this section show that, overall, measures of democracy are convergent and strongly associated among one another. We would expect the four indices to be highly interchangeable, even if we have seen as well that their convergence has mostly decreased in the last decades. It seems that the wave of democratisations that begun in the mid-1970s has not been captured in the same way by the measures.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Table 3 Longitudinal regressions of democracy on structural predictors (1973–2017)

	Democracy			
	PIV	VAN	FH	VDEM
Ln(GDP/capita)	0.151	0.493***	0.342*	0.304***
	(1.110)	(3.748)	(2.200)	(3.392)
GDP/capita growth	0.003	− 0.024	0.016	0.003
	(0.098)	(− 1.103)	(0.676)	(0.207)
Openness	− 0.005*	− 0.003	− 0.004	− 0.004**
	(− 2.123)	(− 1.384)	(− 1.842)	(− 2.829)
Inflation	− 0.00001	− 0.0001	− 0.0001	0.0001
	(− 0.075)	(− 0.636)	(− 1.002)	(1.283)
Education	0.128	0.288***	0.102	− 0.007
	(1.848)	(4.100)	(1.494)	(− 0.218)
Ln(population)	− 0.046	− 0.005	− 0.133	− 0.065
	(− 0.437)	(− 0.052)	(− 1.276)	(− 0.985)
Presidential	0.610	0.732*	0.644	0.683*
	(1.364)	(2.246)	(1.812)	(2.507)
Parliamentary	0.370	1.176***	0.874*	0.525
	(0.895)	(3.749)	(2.437)	(1.830)
Party fractionalisation	1.724**	1.577**	0.533	0.077
	(2.683)	(2.709)	(1.010)	(0.241)
R^2	0.95	0.94	0.96	0.98
Observations	4042	4042	4042	4042
Countries	127	127	127	127

Note: *T*-values in parentheses; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Constant coefficient measured, but not reported. All models are estimated with OLS-PCSE and include a lagged dependent variable (not reported) on the right-hand side of the regression equation. Independent variables are lagged by 1 year. The years 1973–2017 refer to the dependent variable

To sum up, VAN has become considerably less associated with the other measures. Likewise, PIV has become less associated with the other measures, but not as much as VAN. The only exception is the relationship between FH and VDEM, which has become slightly stronger from 1975 to 2005 and remained relatively stable from 2005 on.

Interchangeability

Table 3 presents the results of the first set of longitudinal regressions of measures of democracy on structural predictors. All other things equal, the results show that different measures of democracy can lead to substantially different findings and interpretations. According to this first set of models, frequently used measures of democracy are not interchangeable at all.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

The only similarities are that GDP/capita growth, inflation, and population are not significant predictors of democracy in any of the models. All the other associations depend meaningfully on the chosen measure of democracy. GDP/capita is a positive and significant predictor of democracy only when democracy is quantified with VAN (at the 99.9% level), VDEM (at the 99.9% level), or FH (at the 95% level). On the contrary, when democracy is quantified with PIV, there is no significant relationship between GDP/capita and the level of democracy. Trade openness is significantly and inversely related to democracy with PIV (at the 95% level) and VDEM (at the 99% level), but with VAN and FH the relationship is not significantly different from zero. Education has a significant positive effect on democracy with VAN, whereas with PIV, FH, and VDEM there is no significant relationship between education and democracy. Presidential systems are significantly and positively related to democracy only with VAN and VDEM (in both cases at the 95% level). Parliamentary systems are significantly and positively related to democracy only with FH (at the 95% level) and VAN (at the 99.9% level). There is a significant and positive relationship between party fractionalisation and democracy only with PIV and VAN. With FH and VDEM, instead, the relationship between party fractionalisation and democracy is not even close to conventional levels of statistical significance.

The findings of the first set of regressions show unquestionably that measures of democracy are weakly interchangeable in our full sample of common country-years. Nevertheless, since we previously found that in general the convergence among measures of democracy has become weaker over time, I investigate whether the interchangeability of the measures follows a similar pattern as well. I divide the dataset into two similar parts (1973–1995 [$T=23$]; 1996–2017 [$T=22$]) and regress each measure of democracy on our selection of structural predictors. Due to our earlier findings, we would expect the measures to be more interchangeable between 1973 and 1995 than between 1996 and 2017.

The expectations are confirmed by the regression estimates. In the first sample of years (1973–1995), the estimated results lead to similar interpretations nearly without any exceptions (Table 4). Only the effect of education depends meaningfully on the chosen measure of democracy: education is significantly and positively related to democracy with VAN and FH, but has no significant relationship with democracy with PIV and VDEM. No other prediction is statistically different from zero at conventional levels. This means that, in general, in the first sample of years (1973–1995), the choice of the measure does not play a decisive role in generating different results. In fact, from 1973 to 1995, PIV and VDEM lead exactly to the same conclusions across all predictors. Given these results, it would not be surprising to find that the weak interchangeability of the full sample (1973–2017) is driven by increased divergencies among measures in more recent years (1996–2017).

As suspected, in the second sample of years (1996–2017) the choice of the measure of democracy affects crucially the conclusions (Table 5). With VAN and VDEM, the relationship between GDP/capita and democracy is positive and significant (at the 99% level), whereas with the other two measures of democracy there is no evidence of a significant relationship between GDP/capita and democracy. Trade openness is a significant predictor of democracy only with PIV and VDEM, education is a significant predictor of democracy only with VAN, and population predicts



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Table 4 Longitudinal regressions of democracy on structural predictors (1973–1995)

	Democracy			
	PIV	VAN	FH	VDEM
Ln(GDP/capita)	0.068	0.493	0.425	0.079
	(0.204)	(1.732)	(1.065)	(0.352)
GDP/capita growth	0.014	− 0.041	0.050	0.014
	(0.333)	(− 1.361)	(1.251)	(0.524)
Openness	− 0.002	− 0.002	− 0.005	0.001
	(− 0.330)	(− 0.400)	(− 1.205)	(0.446)
Inflation	0.0001	0.00002	− 0.0002	0.0002
	(0.240)	(0.183)	(− 0.868)	(0.759)
Education	0.316	0.502**	0.393*	0.146
	(1.425)	(3.070)	(2.110)	(1.654)
Ln(population)	0.068	0.041	− 0.278	0.183
	(0.297)	(0.253)	(− 1.172)	(1.436)
Presidential	− 0.093	0.327	0.765	1.223
	(− 0.088)	(0.668)	(0.859)	(1.847)
Parliamentary	− 0.143	0.820	1.464	0.869
	(− 0.170)	(1.254)	(1.736)	(1.550)
Party fractionalisation	1.212	0.941	0.882	− 0.268
	(0.948)	(0.969)	(0.628)	(− 0.474)
R^2	0.94	0.95	0.93	0.97
Observations	1474	1474	1474	1474
Countries	108	108	108	108

Note: T-values in parentheses; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Constant coefficient measured, but not reported. All models are estimated with OLS-PCSE and include a lagged dependent variable (not reported) on the right-hand side of the regression equation. Independent variables are lagged by 1 year. The years 1973–1995 refer to the dependent variable

democracy significantly only with VDEM. The relationship between presidential systems and democracy is significantly positive with PIV (at the 95% level), VAN (at the 99% level), but non-significant with FH and VDEM. Parliamentary systems are positively related to democracy with VAN (at the 99.9% level), whereas with PIV, FH, and VDEM the estimates do not point to the existence of such a relationship. Party fractionalisation has a positive and significant effect on democracy with PIV and VAN, but not with FH and VDEM.

The coefficient plots in Fig. 3 provide a visual representation of the point estimates and 95% confidence intervals of selected independent variables in our battery of regression models. These plots illustrate clearly how the lack of interchangeability in the full sample of country-years (1973–2017, first row in Fig. 3) is driven by the weak interchangeability of the measures in the sample of most recent years (1996–2017, third row in Fig. 3). In the antecedent sample of



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Table 5 Longitudinal regressions of democracy on structural predictors (1996–2017)

	Democracy			
	PIV	VAN	FH	VDEM
Ln(GDP/capita)	0.121	0.477**	0.191	0.311**
	(0.775)	(2.832)	(1.352)	(2.719)
GDP/capita growth	0.005	0.0003	0.001	0.001
	(0.164)	(0.010)	(0.042)	(0.069)
Openness	− 0.006*	− 0.003	− 0.004	− 0.005**
	(− 2.173)	(− 0.884)	(− 1.607)	(− 2.652)
Inflation	− 0.0002	− 0.0002	− 0.0001	− 0.0001
	(− 0.895)	(− 1.880)	(− 0.363)	(− 0.907)
Education	0.077	0.218*	0.028	− 0.023
	(1.206)	(2.550)	(0.389)	(− 0.672)
Ln(population)	− 0.109	− 0.026	− 0.057	− 0.161*
	(− 1.107)	(− 0.242)	(− 0.600)	(-2.342)
Presidential	0.969*	1.003**	0.610	0.395
	(2.165)	(2.688)	(1.729)	(1.388)
Parliamentary	0.603	1.391***	0.577	0.231
	(1.259)	(3.622)	(1.516)	(0.625)
Party fractionalisation	2.236**	2.055*	0.681	0.611
	(3.007)	(2.451)	(1.233)	(1.682)
R^2	0.95	0.94	0.97	0.98
Observations	2568	2568	2568	2568
Countries	127	127	127	127

Note: T-values in parentheses; *$p<0.05$, **$p<0.01$, ***$p<0.001$. Constant coefficient measured, but not reported. All models are estimated with OLS-PCSE and include a lagged dependent variable (not reported) on the right-hand side of the regression equation. Independent variables are lagged by 1 year. The years 1996–2017 refer to the dependent variable

country-years (1973–1995, second row in Fig. 3), on the contrary, the results are almost never different from zero, regardless of the chosen measure.

Besides summarising the main results of the regressions, coefficient plots are particularly useful to analyse differences in confidence intervals. They allow us to notice immediately that, almost without exceptions, the predicted effects in models with VDEM as the dependent variable have systematically narrower standard errors compared to the other models. Being the cause of less precise predictions can be considered a limitation of any measure, since we can confidently assume that users of measures of democracy prefer precise inferences to imprecise inferences. This finding can be explained by the fact that in general VDEM captures much more variation across countries and over years than the other measures. Hence, VDEM seems to lead to more precise estimates than the other measures and is likely to bring more fruitful results in large-N cross-national studies than VAN, PIV, or FH.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

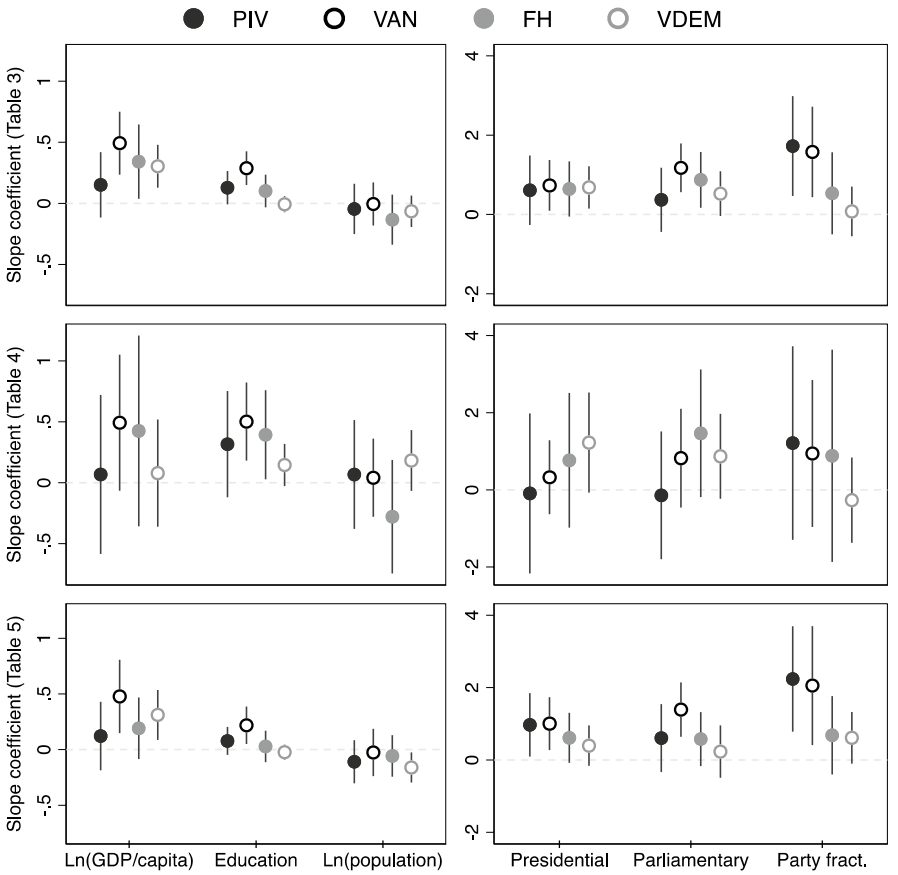


Fig. 3 Coefficient plots of selected predictors across regression models

We would expect highly convergent measures of democracy to be similarly related to structural predictors of democracy, but the findings show instead that in general the interchangeability of measures of democracy is weak. Choosing one measure instead of another can affect substantially the conclusions. In fact, in two of our three sets of regression models there is not even a single pair of measures that leads to the same conclusions across all predictors. This is problematic, since most scholars treat highly correlated measures of democracy as interchangeable. The findings show as well that the interchangeability of these measures has weakened over time. While the measures lead to relatively similar conclusions in a longitudinal analysis from 1973 to 1995, the choice of the measure becomes crucial in a longitudinal analysis from 1996 to 2017, where all measures lead to different conclusions.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Conclusions

Through a comparative statistical analysis of four frequently used democracy indices, the study at hand contributes to the literature on the measurement of democracy and provides novel information regarding the statistical properties, convergence, and interchangeability of some of the most currently relevant measures.

First of all, the analysis of the statistical properties has shown that none of the democracy measures is normally distributed and some of the measures tend to clump at the extremes of the scale. In general, VAN compresses too many observations at the bottom extreme of the scale and PIV and FH compress too many observations at the top extreme of the scale. These three measures could be improved by extending their scales. VDEM, PIV, and FH are more capable than VAN to distinguish between different degrees of democracy at the low end of the scale, whereas VDEM and VAN are more capable than PIV and FH to distinguish between different degrees of democracy at the high end of the scale. There is no agglomeration of observations at the extremes with VDEM.

Regardless of these differences, the findings of the study at hand have shown that frequently used measures of democracy are highly convergent among one another. Anyhow, this convergence has not stayed still over time, but instead, we have observed a pattern of decreasing bivariate association between most measures since the 1980s. While in 1980 and 1985 most bivariate correlation coefficients between the measures of democracy were at least 0.90, in 2015 they were below 0.90 nearly without exceptions. Especially VAN has become less related to the other measures, although it is important to stress that, overall, the surveyed measures of democracy are strongly related to one another throughout the analysed period of time.

Despite high correlations, the findings show that the interchangeability of measures of democracy is generally weak, providing robustness to some of the previous claims on the contrasting relationship between correlation and interchangeability of measures of democracy. By substantially broadening the empirical foundations of earlier studies on the topic with additional measures and more extensive longitudinal coverage, this study provides novel knowledge on the interchangeability of currently relevant measures of democracy. The estimations from three sets of longitudinal regressions show not only that the choice of the measure of democracy affects significantly the conclusions, but also that the interchangeability of these measures has become weaker during the last few decades. On the whole, it is worrying that measures of democracy can be so differently related to common predictors of democracy, but it explains some of the inconsistent conclusions related to democracy, its causes, and its consequences.

If findings are not generalisable across measures, it becomes important to understand why. This comparative analysis has pointed out some statistical differences and shortcomings of the measures that scholars should take into consideration when using these measures. To give a few examples, we have seen that VAN fails to distinguish between highly autocratic countries, that PIV and FH



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

compress countries at the extremes of the scale, and that VDEM captures most of the variation between countries. VDEM seems to "cause" also more precise estimations than the other measures. Moreover, scholars should remember that VAN has become less and less associated with the other measures since 1980, raising some doubts regarding its validity and reliability.

Individual country scores can provide more detailed insights on some of the increased divergencies among the measures. While in the early years of our sample it is hard to find severe disagreements about countries, in the last couple of decades there are disagreements even on the most obvious cases. For instance, virtually all experts would agree that Venezuela experienced a major democratic erosion during Chávez's presidency of the country. However, a comparison of the country-level scores in Venezuela from 1999 to 2013 shows that this "democratic backsliding" is captured only by VDEM, FH, and PIV. Incredibly, according to VAN, the level of democracy in Venezuela increased considerably during Chávez's rule.

Similarly, it is commonly acknowledged that Hungary has suffered a democratic decline in the last years, but this decline is not captured by all our measures of democracy. From 2010 to 2018, the level of democracy in Hungary has decreased sharply according to VDEM and FH, but has remained stable and as high as possible according to PIV and has even slightly increased according to VAN. Since there is little doubt that in practice both Venezuela and Hungary have suffered a decline in democracy, an inability to record such pattern must be considered as a shortcoming. Besides of being aware of the weak interchangeability, scholars using measures of democracy should be aware of these deficiencies in individual country scores. In particular, these two examples confirm our questions regarding the validity of VAN and suggest that comparative studies of individual country scores could be valuable in understanding more thoroughly divergencies among measures of democracy.

To conclude, as we have seen in this study, measures of democracy are not equivalent nor interchangeable despite high convergence. The choice of the measure matters, and ideally, researchers should carefully choose the measure of democracy that best matches their theory and defend the choice on theoretical grounds. When this is the case, the lack of interchangeability becomes a minor problem. If scholars are unwilling to make such a move, however, they should at the minimum test the robustness of their results with several frequently used measures of democracy (at least with VDEM, PIV, and FH). Even more so in studies on the last few decades, since the findings have pointed out that the interchangeability of measures of democracy has decreased. Alternatively, if theoretically appropriate, researchers could dig deeper into the sub-components of the measures and select a specific subcomponent that matches well their theory. Whatever the chosen approach, the study at hand has shown convincingly that since measures of democracy are weakly interchangeable, findings are not likely to be valid across different measures, unless further evidence to generalise findings is provided.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1057/s41304-021-00328-8.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alvarez, M., J.A. Cheibub, F. Limongi, and A. Przeworski. 1996. Classifying Political Regimes. *Studies in Comparative International Development* 31 (2): 3–36.

Anckar, C., and C. Fredriksson. 2019. Classifying Political Regimes 1800–2016: A Typology and a New Dataset. *European Political Science* 18 (1): 84–96.

Barsh, R.L. 1993. Measuring Human Rights: Problems of Methodology and Purpose. *Human Rights Quarterly* 15 (1): 87–121.

Beck, N., and J.N. Katz. 1995. What to Do (and Not to Do) with Time-Series Cross-Section Data. *The American Political Science Review* 89 (3): 634–647.

Besançon, M. 2003. *Good Governance Rankings: The Art of Measurement*. Cambridge, MA: World Peace Foundation.

Boese, V.A. 2019. How (Not) to Measure Democracy. *International Area Studies Review* 22 (2): 95–127.

Bollen, K.A. 1980. Issues in the Comparative Measurement of Political Democracy. *American Sociological Review* 45 (3): 370–390.

Bollen, K.A., and P. Paxton. 2000. Subjective Measures of Liberal Democracy. *Comparative Political Studies* 33 (1): 58–86.

Bowman, K., F. Lehoucq, and J. Mahoney. 2005. Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America. *Comparative Political Studies* 38 (8): 939–970.

Casper, G., and C. Tufis. 2003. Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets. *Political Analysis* 11 (2): 196–203.

Coppedge, M., A. Alvarez, and C. Maldonado. 2008. Two Persistent Dimensions of Democracy: Contestation and Inclusiveness. *Journal of Politics* 70 (3): 632–647.

Coppedge, M., J. Gerring, C.H. Knutsen, S.I. Lindberg, J. Teorell, D. Altman, M. Bernhard, M.S. Fish, A. Glynn, A. Hicken, A. Lührmann, K.L. Marquardt, K. McMann, P. Paxton, D. Pemstein, B. Seim, R. Sigman, S.E. Skaaning, J. Staton, A. Cornell, L. Gastaldi, H. Gjerløw, N. Ilchenko, J. Krusell, L. Maxwell, V. Mechkova, J. Medzihorsky, J. Pernes, J. von Römer, N. Stepanova, A. Sundtröm, E. Tzelgov, Y. Wang, T. Wig, and D. Ziblatt. 2019. *V-Dem [Country-Year/Country-Date] Dataset v9*. Varieties of Democracy Project. https://doi.org/10.23696/vdemcy19.

Coppedge, M., J. Gerring, S.I. Lindberg, S.E. Skaaning, and J. Teorell. 2017. V-Dem Comparisons and Contrasts with Other Measurement Projects. *The Varieties of Democracy Institute Working Paper Series* 2017: 45.

Dahl, R.A. 1971. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Eckstein, H., and T.R. Gurr. 1975. *Patterns of Authority: A Structural Basis for Political Inquiry*. New York: Wiley.

Economist Intelligence Unit. 2019. *Democracy Index 2019: A Year of Democratic Setbacks and Popular Protest*. London: Economist Intelligence Unit.

Elff, M., and S. Ziaja. 2018. Method Factors in Democracy Indicators. *Politics and Governance* 6 (1): 92–104.

Elkins, Z. 2000. Gradations of Democracy? Empirical Tests of Alternative Conceptualizations. *American Journal of Political Science* 44 (2): 293–300.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Freedom House. 2019a. *Freedom in the World 2019 Data*. https://freedomhouse.org/report-types/freedom-world.
Freedom House. 2019b. *Methodology 2019*. https://freedomhouse.org/report/methodology-freedom-world-2019.
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.
Gleditsch, K.S., and M.D. Ward. 1997. Double Take: A Reexamination of Democracy and Autocracy in Modern Polities. *Journal of Conflict Resolution* 41 (3): 361–383.
Goertz, G. 2020. *Social Science Concepts and Measurement: New and Completely*. Revised. Princeton: Princeton University Press.
Gurr, T.R. 1974. Persistence and Change in Political Systems, 1800–1971. *American Political Science Review* 68 (4): 1482–1504.
Hadenius, A., and J. Teorell. 2005. Assessing Alternative Indices of Democracy. *IPSA Committee on Concepts and Methods Working Papers* 6 (August 2005). http://www.concepts-methods.org/WorkingPapers/PoliticalConcepts.
Henisz, W.J. 2017. *The Political Constraint Index (POLCON) Dataset 2017 Data Release*. https://mgmt.wharton.upenn.edu/faculty/heniszpolcon/polcondataset.
Hintze, J.L., and R.D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism Statistical Computing and Graphics. *The American Statistician* 52 (2): 181–184.
Högström, J. 2013. Does the Choice of Democracy Measure Matter? Comparisons between the Two Leading Democracy Indices, Freedom House and Polity IV. *Government and Opposition* 48 (2): 201–221.
Kaufmann, D., A. Kraay, and M. Mastruzzi. 2011. The Worldwide Governance Indicators: Methodology and Analytical Issues. *Hague Journal on the Rule of Law* 3 (2): 220–246.
Knutsen, C.H. 2010. Measuring Effective Democracy. *International Political Science Review* 31 (2): 109–128.
Marshall, M.G., T.R. Gurr, and K. Jaggers. 2019. *Polity IV Project Data [p4v2018], Political Regime Characteristics and Transitions, 1800–2018*. Center for Systemic Peace. https://www.systemicpeace.org/inscrdata.html.
McMahon, E., and E. Kornheiser. 2010. Assessing the Assessors: Correlating Democracy Methodologies. *Social Indicators Research* 97 (2): 269–277.
Møller, J., and S.E. Skaaning. 2011. On the Limited Interchangeability of Rule of Law Measures. *European Political Science Review* 3 (3): 371–394.
Munck, G.L., and J. Verkuilen. 2002. Conceptualizing and Measuring Democracy. *Comparative Political Studies* 35 (1): 5–34.
Pemstein, D., S.A. Meserve, and J. Melton. 2010. Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. *Political Analysis* 18 (4): 426–449.
Pickel, S., T. Stark, and W. Breustedt. 2015. Assessing the Quality of Quality Measures of Democracy: A Theoretical Framework and Its Empirical Application. *European Political Science* 14 (4): 496–520.
Seawright, J., and D. Collier. 2014. Rival Strategies of Validation: Tools for Evaluating Measures of Democracy. *Comparative Political Studies* 47 (1): 111–138.
Skaaning, S.E., J. Gerring, and H. Bartusevičius. 2015. A Lexical Index of Electoral Democracy. *Comparative Political Studies* 48 (12): 1491–1525.
Teorell, J., M. Coppedge, S. Lindberg, and S.E. Skaaning. 2019. Measuring Polyarchy Across the Globe, 1900–2017. *Studies in Comparative International Development* 54 (1): 71–95.
Treier, S., and S. Jackman. 2008. Democracy as a Latent Variable. *American Journal of Political Science* 52 (1): 201–217.
Vanhanen, T. 2000. A New Dataset for Measuring Democracy, 1810–1998. *Journal of Peace Research* 37 (2): 251–265.
Vanhanen, T. 2019. *Measures of Democracy 1810–2018 Version 8.0*. Finnish Social Science Data Archive. http://urn.fi/urn:nbn:fi:fsd:T-FSD1289.
World Bank. 2017. *World Development Indicators [December 2017]*. The World Bank. https://datatopics.worldbank.org/world-development-indicators/.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Andrea Vaccaro is a PhD student at the Department of Social Sciences and Economics at the Sapienza University of Rome. Most recently, he has been working on the measurement of democracy and state capacity. More broadly, in addition to democracy, state capacity, and the relationship between the two, his research interests include socio-economic development, measurement issues, and quantitative methods.



https://doi.org/10.1057/s41304-021-00328-8 Published online by Cambridge University Press