Advance Access publication August 26, 2010

Political Analysis (2010) 18:426–449
doi:10.1093/pan/mpq020

Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type

Daniel Pemstein
Center for the Study of Democratic Institutions, Vanderbilt University, PMB 407712, 2301 Vanderbilt Place, Nashville, TN 37240-7712
e-mail: dan.pemstein@vanderbilt.edu (corresponding author)

Stephen A. Meserve
Department of Political Science, University of Illinois at Urbana-Champaign, 240 Computer Applications Building, 605 East Springfield Ave., Champaign, IL 61820
e-mail: meserve@illinois.edu

James Melton
Economics and Institutional Change, IMT Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca, Italy
e-mail: james.melton@imtlucca.it

Using a Bayesian latent variable approach, we synthesize a new measure of democracy, the Unified Democracy Scores (UDS), from 10 extant scales. Our measure eschews the difficult—and often arbitrary—decision to use one existing democracy scale over another in favor of a cumulative approach that allows us to simultaneously leverage the measurement efforts of numerous scholars. The result of this cumulative approach is a measure of democracy that, for every country-year, is at least as reliable as the most reliable component measure and is accompanied by quantitative estimates of uncertainty in the level of democracy. Moreover, for those who wish to continue using previously existing scales or to evaluate research performed using those scales, we extract information from the new measure to perform heretofore impossible direct comparisons between component scales. Specifically, we estimate the relative reliability of the constituent indicators, compare the specific ordinal levels of each of the existing measures in relationship to one another and assess overall levels of disagreement across raters. We make the UDS and associated parameter estimates freely available online and provide a detailed tutorial that demonstrates how to best use the UDS in applied work.

Authors' Note: The authors are listed in reverse alphabetical order, indicating equal contribution to the article. The authors would like to thank Bill Bernhard, Jose Antonio Cheibub, Josh Clinton, Zach Elkins, Brian Gaines, Jim Kuklinski, Kevin Quinn, seminar participants at the University of Illinois, participants at the 24th Annual Meeting of the Society for Political Methodology, our reviewers, and the editors of Political Analysis for helpful comments. Supplementary materials for this article are available on the *Political Analysis* Web site.

1 Introduction

Democracy is a fundamental concept of politics. Yet, like many constructs in the social sciences, it is unobservable and poses a difficult measurement problem for quantitative analysts. Nonetheless, there has been no lack of enthusiasm among researchers when it comes to measuring democracy. Indeed, democracy scales proliferate, and Munck and Verkuilen (2002) identify nine well-known measures of democracy in their survey of the field. This bewildering array of options confronts applied researchers with a dilemma: what metric provides the most valid and reliable measure of democracy and which indicator is most suitable for use in any particular application?

The literature lacks a concrete answer to these questions. The available measures correlate highly with one another and generally appear to tap the same underlying concept (Adcock and Collier 2001). This is somewhat surprising considering the wide array of conceptualization, measurement, and aggregation techniques employed by the creators of these scales (Munck and Verkuilen 2002). However, despite the similarities in these scales, the subtle differences between them can and do affect substantive results (Elkins 2000; Casper and Tufis 2002). Moreover, although few democracy raters quantify the uncertainty around their point estimates, recent research has shown that the error variability of popular measures is large enough to render all but the most dissimilar of regimes statistically indistinguishable, undermining scholars' ability to confidently employ existing scales in applied research (Treier and Jackman 2008).[1] Even dismissing these two problems, the pragmatic recommendation to select the single measure that best operationalizes democracy for the question at hand (Collier and Adcock 1999) forces researchers to throw out potentially useful information embedded in other available democracy scales.

To help alleviate these problems, this paper eschews the difficult—and often arbitrary—decision to use one particular democracy scale over another and favors a cumulative approach that allows us to leverage the measurement efforts of numerous scholars simultaneously. Building on a model of the democracy rating process, we synthesize a new measure of democracy from existing scales, the Unified Democracy Scores (UDS). The UDS average over the uncertainty inherent in each of the constituent measures, taking advantage of each scale's tendency to capture similar, but often distinct, aspects of what makes states more or less democratic. Furthermore, we accompany this new scale of democracy with quantitative estimates of measurement error and demonstrate that, by exploiting the combined efforts of other researchers, we are able to significantly improve confidence in estimates beyond what is possible using only a single measure.

The UDS do not simply improve measurement confidence but also minimize the impact of idiosyncratic errors that occur in individual measures and take advantage of the level of agreement between raters to perform a form of intercoder validation across major democracy scales. The UDS are also flexible and incorporate information both from measures spanning a handful of country-years and multiyear global projects. Perhaps, most importantly, the UDS allow scholars to effectively leverage the immense effort that other researchers have invested in creating democracy scores. Using a democracy scale in one's research will no longer force scholars to make an arbitrary choice and casually cast aside the vast majority of the information available on the topic. This is especially important in situations where extant scales provide divergent estimates of democracy level; the

[1] Although not all the democracy measures we discuss in this article are ratings in the strictest sense, we adopt the term "raters" to describe the producers of democracy scores, in keeping with standard terminology in the statistical literature on multiple measures.

confidence intervals around the UDS reflect these disagreements and allow researchers to deal with these cases in a reasoned and systematic manner. Our approach also makes possible the direct analysis of various scales in relation to one another, helping researchers to calibrate cutpoints across measures and easily compare substantive results arrived at with different scales. Finally, the model provides estimates of rater reliability, generating useful criteria on which to judge the relative performance of the existing democracy scales and allows researchers to rigorously examine differences across extant scales.

2 A Plethora of Measures

The UDS incorporate information from 10 measures of democracy: Arat (1991), Bowman, Lehoucq, and Mahoney (2005) (BLM), Bollen (2001), Freedom House (2007), Hadenius (1992), Przeworski et al. (2000) (PACL), Polity scores by Marshall, Jaggers, and Gurr (2006), Polyarchy scale by Coppedge and Reinicke (1991), Gasiorowski's (1996) Political Regime Change measure (PRC), and Vanhanen (2003).[2] All 10 measures are based on similar underlying conceptualizations of democracy. Munck and Verkuilen (2002, 9), discussing 9 out of the 10 measures, note that ". . . the decision to draw, if to different degrees, on Dahl's (1972, 4–6) influential insight that democracy consists of two attributes—contestation or competition and participation or inclusion—has done much to ensure that these measures of democracy are squarely focused on theoretically relevant issues." But each measure brings different strengths and weaknesses to the table. The most popular measures, such as Freedom House, PACL, and Polity, provide extensive spatial and temporal coverage but may sacrifice a degree of case familiarity. Other measures, like BLM, provide limited coverage but are based on in-depth analyses of primary sources. The raters choose to incorporate different characteristics—subjective and objective—in their scores and use varying techniques to aggregate components. Thus, each judge operationalizes Dahl's (1972) conceptualization of democracy differently and provides potentially valuable information not available in other scores, as Table 1 summarizes.

Notwithstanding their differences, does it really matter which measure scholars' use in their research? Both large-N studies (Elkins 2000; Casper and Tufis 2002) and case evidence suggest that it can. Figure 1, which displays rescaled Freedom House, PACL, and Polity scores for Spain, Russia, Fiji, and Burundi, can help us explore this question in more detail.[3] In general, the available democracy measures correlate highly, and this fact is often used as evidence of the (convergent) validity of the measures (Bollen 1980; Adcock and Collier 2001). The example of Spain in figure 1 underscores the general agreement between these measures for most country-years. The three highlighted raters generally agreed that Spain was an authoritarian regime until around 1975 when Franco died, after which they all scored Spain as largely democratic. Spain is an excellent example of the overall face validity demonstrated by these measures, reflecting the convergence in democracy measures across the majority of the country-years in the data set.

By contrast, both Russia and Fiji highlight disagreement between measures, but the disagreement appears to come from two different sources. In Russia, the 1996 presidential election and the 1998 financial crisis highlighted both the power of the oligarchs in Russian politics and the precarious relationship between President Yeltsin and the Duma. As a result of these two events, Freedom House lowered its rating of Russia's level of democracy,

[2] We use the extended version of Przeworski et al. (2000) data set compiled by Cheibub and Gandhi (2010). Similarly, we use Reich's (2002) extension to the PRC data.

[3] We rescaled each rater's score in figure 1 to the (0, 1) interval.

Table 1 Ten measures of democracy

Measure	*Countries*	*Years*	*Scale*	*Components*
Arat	65–150	1948–1982	29–109	Participation, Inclusiveness, Competitiveness, and Coerciveness
BLM	5	1946–2000	0.0, 0.5, or 1.0	Political Liberties, Competitive Elections, Inclusive Participation, Civilian Supremacy, and National Sovereignty
Bollen	60, 70, 105, 117, and 158	1950, 1955, 1960, 1965, and 1980	0–100	Political Liberties and Popular Sovereignty
Freedom House	135–191	1972–2000	1–7	Political Rights and Civil Liberties
Hadenius	129	1988	0–10	Elections and Political Freedoms
PACL	66–189	1946–2000	0 or 1	Executive Elections, Legislative Elections, and Party Competition
Polity	60–151	1946–2000	−10 to –10	Competitiveness of Participation, Regulation of Participation, Competitiveness of Executive Recruitment, Openness of Executive Recruitment, and Constraints on the Executive
Polyarchy	162 and 191	1985 and 2000	0–10	Free and Fair Elections, Freedom of Organization, Freedom of Expression, and Pluralism in the Media
PRC	64–143	1946–1998	1–4	Competitiveness, Inclusiveness, and Political Liberties
Vanhanen	41–155	1946–2000	0.01–53.81	Competition and Participation

Note. The measures were drawn from the following sources: Arat (Arat 1991), BLM (Bowman, Lehoucq, and Mahoney 2005), Bollen (Bollen 2001), Freedom House (Freedom House 2007), Hadenius (Hadenius 1992), PACL (Przeworski et al. 2000; Cheibub and Gandhi 2010), Polity (Marshall, Jaggers, and Gurr 2006), Polyarchy (Coppedge and Reinicke 1991), PRC (Gasiorowski 1996; Reich 2002), and Vanhanen (Vanhanen 2003).

Polity increased its rating of Russia's level of democracy, and PACL's rating remained constant. All three measures had reason for their scores. Russia's score on the Polity scale almost certainly increased as a result of an increase in the perceived strength of the legislature after the Duma's rejection of Yeltsin's nomination for Prime Minister in late 1998 since the score rose as a result of an increase in Polity's executive constraint subcomponent. As for Freedom House, one can speculate that the role of the oligarchs, both during the 1996 election and 1998 financial crisis, hurt Russia's rating on their scale. Russia seems to be a case where two raters looked at the same information and came to different conclusions. Both Freedom House and Polity made judgments that were sensible estimates of the level of democracy in Russia, given their measurement strategies, but each estimate was incomplete. In cases like this, choosing a score invariably involves sacrificing relevant information. In such circumstances, the UDS provide a sensible alternative, weighing the contribution of each score in terms of its overall reliability.

Fiji also highlights a major disagreement between measures, but the disagreement is likely the result of a lack of information, rather than divergence in raters' informational focus. PACL consistently ranks Fiji as an authoritarian regime based on their type II error

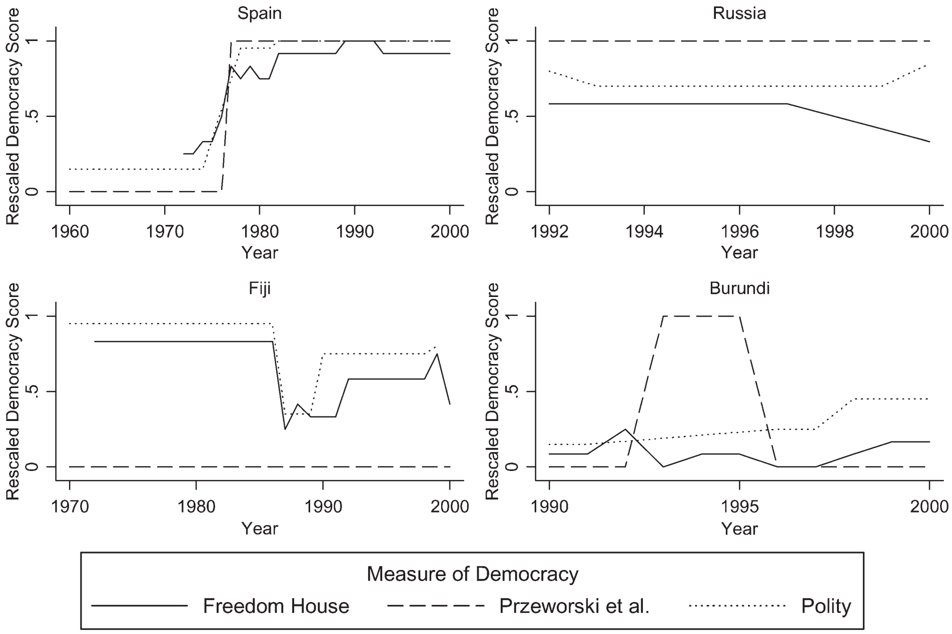

Fig. 1 Rescaled measures of democracy over time.

rule. In other words, because the Fijian Alliance won every election until 1987,[4] there was no way to know if alternation in office would have occurred if the Indo-Fijians won, so PACL judges it better to classify Fiji as an authoritarian regime during the period from 1970 to 1987. Polity and Freedom House, on the other hand, both look at the Fijian government from 1970 to 1987 as very democratic, despite the lack of alternation in office. The Fijian Alliance had fairly won each election, indicating a democratic regime by Polity and Freedom House standards. Raters simply do not have enough information to classify the Fijian regime from 1970 to 1987 because they could not observe a counterfactual Indo-Fijian victory. Unable to arbitrate definitively between measures, we believe that a more honest assessment of democracy level would take the opinions of multiple raters into account.

The final country in figure 1 is Burundi. Burundi is an example where none of these three measures can agree on a score between 1993 and 1996. PACL scores the Burundian regime during this time as democratic, Polity scores it as transitional, and Freedom House scores it as authoritarian. In some sense, each of these raters is correct. The elections and Hutu succession in office in 1994 buttress PACL, but the instability in government and continual fight for power by both the Hutus and the Tutsis support a more transitional view, similar to Polity. However, the thousands killed by both sides during the civil war and repression of human rights by both the Hutus and Tutsis support Freedom House. This is a classic example of uncertainty in measurement. In this case, simple point estimates—from any scale—do not fully capture our knowledge and beliefs about the level of democracy in Burundi.

[4] The Indo-Fijians did win an election in 1978, but since they could not form a government, the governor-general called new elections, which were won by the Fijian Alliance.

Any scholar working with measurements like these should temper her conclusions by incorporating some estimate of democracy score confidence into the analysis, something the UDS make possible.

Just how common are the major discrepancies between measures? It is hard to tell when simply looking at raw scores, but a number of countries demonstrate each of the patterns illustrated in figure 1. We address this question further in Section 6.3, where we find that at least two raters' scores are statistically distinguishable in as many as 50% of the country-years we consider. The implications for scholars using only a single measure in analyses are clear: they make all the mistakes of their chosen scale, even when its ratings are at odds with the majority of other raters. We argue that sticking with one measure of democracy, even one crafted by a relatively reliable rater, represents a missed opportunity to utilize the community of democratic scholars' hard work and diverse approaches to operationalizing the concept.

3 Unifying Democracy Measurement

In a sense, the focus of this paper is not on measuring democracy, it is on modeling how other researchers rate or judge democracy across polities. Nonetheless, it is impossible to describe the behavior of raters without first choosing a specific operationalization of the concept of democracy. In the process, we approach serious debates in the literature on democratic measurement in a pragmatic manner. A fundamental question about democracy is whether it is a graded concept or a dichotomous one (Collier and Adcock 1999). Yet, of the 10 measures, we consider here all, but one provide ordinal or continuous estimates. Thus, although the field is divided on this topic, it makes sense to bend to the will of the majority when evaluating existing indices. Similarly, although scholars have argued that democracy is a multidimensional entity (Coppedge, 2002; Munck and Verkuilen 2002; Coppedge, Alvarez, and Maldonado 2008), the raters we examine here all provide a single summary value for each observation. And, although some of these judges also publish scores for various dimensions or subcomponents of democracy, democracy measures are generally used as simple more-or-less ratings in applied work. Therefore, following Bollen and Jackman (1989) and Treier and Jackman (2008), we model each indicator as an approximation to an unobserved—or latent—continuous unidimensional variable.[5]

Specifically, each of the $j = 1, \ldots, m$ judges provides a rating t_{ij} of the level of democracy in each of $i = 1, \ldots,$ country-years. We assume that these ratings roughly capture the true latent level of democracy in each country-year but that raters make mistakes. Therefore, given the true level of democracy z_i, in country-year i, rater j generates a perception t_{ij} of democracy in that country-year, such that

$$t_{ij} = z_i + \mathrm{e}_{ij} \sim N\left(0, \sigma_j^2\right) \tag{1}$$

or, in other words, judge j perceives the true level of democracy accurately on average but makes stochastic mistakes based on her own personal error variance, σ_j^2. Assuming that democracy raters' mistakes are completely nonsystematic is clearly an oversimplification,

[5] The dichotomous measure of Alvarez et al. (1996) is clearly at odds with the continuity assumption. Furthermore, its creators are strong proponents of an either-or approach to conceptualizing democracy. Nonetheless, we believe it is useful to compare the measure of Alvarez et al. (1996) to the graded scales on their terms, and as we will argue, the dichotomous indicator behaves in a manner that is consistent with the idea that it represents a continuous underlying concept.

but this assumption provides a useful starting point for modeling the measurement process. Although complemented by other research (see, e.g., Bollen and Paxton 2000), this model provides a parsimonious base on which to build future work that directly accounts for systematic biases in measures. Furthermore, the differences exhibited by the democracy measures are, largely, a function of a multitude of small effects generated by subtle differences in conceptualization, aggregation, and measurement across raters and by simple coder mistakes—a data generating process (DGP) that is largely consistent with normally distributed random error. Similarly, as Bowman, Lehoucq, and Mahoney (2005) argue, raters must often rely on fragmentary evidence from secondary sources when constructing large panels of democracy scores. Because the realities that determine true democracy level are often idiosyncratic and case specific, all judges are likely to overlook important, but often differing, details when compiling their scores, and it may be difficult to discern systematic relationships between information loss and rater methodology. Thus, although it is often possible to identify the particular ‘‘bias’’ guiding a rater’s judgment in a given case, equation (1) nonetheless represents a reasonable, if imperfect, model of the overall pattern of ratings.[6]

In addition, in so far as the true DGP approximates equation (1), this model provides compelling motivation for integrating the efforts of multiple democracy raters into a single measure. The quantity of interest in this business is z_i, the true level of democracy in country-year i. For the moment, assume that we can directly observe both the raters’ perceptions, t_{ij}, and their error variances, σ_j^2.[7] Suppose we have little information about the value z_i beyond our raters’ perceptions, a situation we can represent by assuming, *a priori*, that z_i is distributed normally with mean zero and some variance σ_0^2.[8] Together, equation (1) and this prior imply that, conditional on t_{ij} and σ_j, z_i has a normal posterior distribution with mean

$$\frac{\sum_{j=1}^{m} \frac{t_{ij}}{\sigma_j^2}}{\frac{1}{\sigma_0^2} + \sum_{j=1}^{m} \frac{1}{\sigma_j^2}} \tag{2}$$

and variance

$$\frac{1}{\frac{1}{\sigma_0^2} + \sum_{j=1}^{m} \frac{1}{\sigma_j^2}}. \tag{3}$$

We learn two things from these equations. First, equation (2) indicates that a mathematically sensible estimate of z_i is simply a weighted average of the prior mean and the individual judges’ perceptions, with weights proportional to individual precisions. Thus, our basic model incorporates information from every available rater but discounts the contributions of less reliable judges. In the statistical model that we describe below, we estimate these rater error variances—each σ_j^2—directly from the data, allowing us to empirically

[6] We assess the plausibility of this assumption when discussing the fit of the model in Section 4. To preview our results, the model fits the data exceedingly well. Therefore, although the component measures may exhibit systematic biases, these biases are not large enough to substantively affect our model’s ability to capture the DGP underlying democracy measurement. Nevertheless, modeling systematic rater error is a promising avenue for future research.

[7] In the real world, we directly observe neither t_{ij} nor σ_j^2. Later, we will build on equation (1) to develop a statistical model that helps us to overcome our observational shortcomings; here, we treat equation (1) as a theoretical model of the rating process and use it to generate some results that motivate our approach to aggregating multiple ratings into a unified set of scores.

[8] When dealing with actual estimation, we assume $\sigma_0^2 = 1$, as we describe later in this section.

determine which judges should most heavily influence the final UDS scores. Furthermore, equation (3) shows that our uncertainty about z_i is decreasing in the number of raters. Each additional rater reduces the variance of the posterior distribution below what is possible using information from any subset of judges, although, clearly, precise raters provide more information about the true level of democracy than unreliable judges, just as was previously reflected in equation (2). This is a—perhaps overly formal—way to hammer home the point that using all the available information and fully capitalizing on the efforts of other researchers can reduce the uncertainty around our democracy estimates.

Although equation (1) lays out a basic way to conceptualize the relationships between democracy ratings, we need to introduce further statistical machinery to deal with various aspects of the reported scales. Fundamentally, we do not observe any of the quantities—z_i, t_{ij}, and σ_j^2—that feature in equation (1). To overcome this problem, we use a technique, multirater ordinal probit, originally developed to compare the performance of multiple essay graders (Johnson 1996; Johnson and Albert 1999).[9] The first obstacle we face is that raters do not report their perceptions—the t_{ij} in equation (1)—directly but rather provide rankings that are based on some unknown function of their underlying perceptions. What we do observe is an n by m matrix of rater scores, $\mathbf{y}$, where y_{ij} describes judge j's reported ranking of country-year i. Six of our ten raters provide ordinal rankings. The four remaining judges—Arat (1991), Bollen (1980), Hadenius (1992), and Vanhanen (2003)—report continuous, ostensibly interval-level, scores.

There are multiple possible ways to parameterize these judges' ratings. One approach to incorporate these four raters into the analysis would be to treat their scores as genuinely interval level and assume some linear relationship between continuous rater j's perception of democracy in country-year i, t_{ij}, and her reported score, y_{ij}, yielding a hybrid model similar to Quinn's (2004) mixed Bayesian factor analysis technique. On the other hand, although these scores take on many values and thus resemble interval scales, they do not necessarily provide interval-level information about democracy levels. In the online appendix, we argue that modeling the continuous measures as ordinal, rather than interval, represents a more conservative—and more empirically valid—approach. Thus, we treat the continuous ratings as ordinal rankings using cutoffs falling at regular intervals along each continuous measure's native scale.[10] Of course, the information we sacrifice when collapsing these measures into a manageable number of rankings might influence parameter estimates and, in turn, the conclusions we draw from the fitted model. To assess this possibility, we estimated the model with a number of different cutoff specifications, and in Section 4, we demonstrate that model estimates are robust to the exact choice of cutoff levels.

After converting the four continuous scales to ordinal rankings, we can treat all 10 scales similarly, and each rater j places each rated country-year i into one of K_j ordered categories, yielding the observed y_{ij}. The scales do not all use the same number of categories, and

[9] The notation that follows borrows liberally from Johnson and Albert (1999). For previous political science applications of similar multirater models, see Jackman (2004), Clinton and Lapinski (2006), and Clinton and Lewis (2007).

[10] We use the following cutpoints: Arat: 50–100, by 10s; Bollen: 10–90, by 10s; Hadenius: 1, 2, 3, 4, 7, 8, 9; Vanhanen: 5–35, by 5s. We are forced to jump between cuts at Hadenius scores of 4 and 7 because of a dearth of observations at levels 5 and 6. Ideally, we would estimate cutpoints at every observed ranking along the continuous measures' scales. Of course, very few rankings fall at exactly the same points along these scales, forcing us to lump ranges together before proceeding with estimation. In sum, we must sacrifice some information and reduce the number of levels in the continuous scores in order to have enough observations at each level to identify cutpoints.

indeed, the measures we examine vary substantially in category-count, ranging between 2 and 22 levels. Furthermore, the model allows for differences in coverage across indicators and C_i is the set of judges who provide a rating for country-year i. The ability to incorporate data sets of varying breadth is extremely useful when dealing with democracy scores and allows us to include information not only from high profile measurement projects with sweeping spatial and longitudinal coverage but also from area experts who, by restricting their sample, are often able to provide highly reliable ratings of a small set of country-years.

To link each observed rating, y_{ij}, to the latent variables, t_{ij} and z_{ij}, introduced in equation (1), we assume that judge j places country-year i in category c if $\gamma_{j,c-1} < t_{ij} \leqslant \gamma_{j,c}$, where $\gamma_{j,c-1}$ and $\gamma_{j,c}$ are judge-specific ranking cutoff points. We fix each $\gamma_{j,0} = -\infty$ and $\gamma_{j,K_j} = \infty$ and $\boldsymbol{\gamma}_{\mathbf{j}} = (\gamma_{j,1}, \ldots, \gamma_{j,K_j-1})$ is the vector of ranking cutoffs for judge j. Thus, for country-year i and rater j, we observe the rating $y_{ij} = c$ if the rater's underlying continuous perception, t_{ij}, of i's true democracy level, falls between $\gamma_{j,\,c-1}$ and $\gamma_{j,\,c}$. Taken together, these assumptions and equation (1) imply the following DGP for our observed rating matrix $\mathbf{y}$:

$$p(y_{ij} = c|z_i, \gamma_j, \sigma_j) = \Phi\left(\frac{\gamma_{j,c} - z_i}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,c-1} - z_i}{\sigma_j}\right), \tag{4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. In other words, given the random variables z_i, γ_j, and σ_j, the probability that judge j places country-year i into category c is the probability that the rater's perception $t_{ij} < \gamma_{j,\,c}$ minus the probability that $t_{ij} < \gamma_{j,\,c-1}$. Multiplying across cases and raters, the likelihood for the observed data matrix $\mathbf{y}$ is

$$L(\mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \prod_{i=1}^{n}\prod_{j \in C_i}\left[\Phi\left(\frac{\gamma_{j,y_{ij}} - z_i}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,y_{ij}-1} - z_i}{\sigma_j}\right)\right]. \tag{5}$$

We can use equation (5) to estimate $\mathbf{z}$, $\boldsymbol{\gamma}$, $\boldsymbol{\sigma}$, and, using an augmented form of equation (5), even $\mathbf{t}$ from the observed data $\mathbf{y}$. Following Johnson and Albert (1999), we identify the model using a Bayesian estimation approach and adopt proper prior distributions for $\mathbf{z}$ and $\boldsymbol{\sigma}$. Specifically, we assume independent standard normal prior distributions for each latent trait z_i—note that this is equivalent to assuming $\sigma_0^2 = 1$ in equations (2) and (3)—and independent inverse-gamma prior densities for each rater variance parameter σ_j^2. Finally, we assume independent uniform priors for each vector of cutoffs γ_j, subject to the constraint that they maintain their order, $\gamma_{j,1} \leqslant \ldots \leqslant \gamma_{j,K_j-1}$. We estimate the model using the Markov chain Monte Carlo (MCMC) algorithm described in Johnson and Albert (1999, 166). We ran the algorithm for 1 million iterations, using the first half of the run as a "burn-in" period and storing every hundredth observation from the second half of the run to create a 5000 observation sample from the join posterior distribution of the model parameters. Standard MCMC diagnostics for the sample are consistent with Markov chain convergence.[11]

The model generates estimates of the latent continuous level of democracy z_i in each polity that are based on the pattern of agreement between the component indicators. Just as

[11] We do not reproduce the conditional posterior distributions of the model parameters or the hybrid Gibbs/Metropolis-Hastings algorithm here because Johnson and Albert (1999) do both, in great detail, for the interested reader. Note, nonetheless, that each country-year's latent democracy level (each z_i) has a Gaussian conditional posterior distribution with mean and variance described by equations (2) and (3), keeping in mind that we assume $\sigma_0^2 = 1$. The online appendix contains further estimation details, including prior parameter values and convergence diagnostics.

in equation (2), these estimates average across individual judges' contributions, weighting each score by its error variance. Furthermore, we can provide probability distributions over these estimates, quantifying the error in the measure and allowing us to evaluate our ability to distinguish between democracies using this approach. In addition, this model lets us estimate the conditional posterior distributions of the judge variance parameters (σ_j^2). These estimates describe the relative precision of the various democracy scales. They are an excellent tool for evaluating the reliability of various measures of democracy and can assist both applied researchers selecting a democracy measure and scholars interested in improving existing indicators or even in building new ones. The multirater model also generates estimates of the $\gamma_{j,}$ rating cutoff point parameters, with confidence intervals. These cutoffs are all scaled to the same underlying latent variable and allow us to compare the scales employed by our judges; for example, these estimates allow us to describe the range of Polity IV scores consistent with an "is-democracy" rating on Alvarez et al. (1996) dichotomy. These cutoff estimates provide a tool with which one can rigorously investigate what scores on the various measures mean in relationship to one another and can improve the comparability of results published by researchers using different democracy scales. Finally, the model fitting algorithm generates estimates of rater perceptions t_{ij}, using the fact that, conditional on the other model parameters, one can sample each t_{ij} from a normal density, truncated to the range $(\gamma_{j,y_{ij}-1}, \gamma_{j,y_{ij}})$, with mean z_i and variance σ_j^2. These estimates are useful for exploring differences across measures, as we demonstrate later.

4 Model Fit and Sensitivity

The model we use to generate the UDS makes strong assumptions about the DGP driving the production of democracy ratings. Most notably, our assumption that raters perceive democracy levels in a noisy but unbiased fashion is, admittedly, quite strong. Although we cannot directly test this assumption, we can examine how well the model fits the data and evaluate the likelihood that the set of ratings we observe could have come from the DGP assumed by our approach. We use posterior predictive checks (Gelman et al., 2004, pp. 159–177) to evaluate the fit of the model to the data. Specifically, we use the fitted model to generate a sample of hypothetical data sets—the posterior predictive distribution (PPD)—from the DGP described by equation (4). We then compare the pattern of scores in the observed data to the PPD to see whether or not the realized data looks like a representative draw from DGP implied by the fitted model.

We first evaluated the model's predictive accuracy on a score-by-score basis by comparing every observed rating with the PPD of ratings for the given case. Overall, 92% of observed ratings fall within the PPD's interquartile ranges, and 99% fall within the PPD's 95% credible intervals, indicating a good match between model and data. There is some variation in model accuracy across raters, but even the most misclassified rater, Vanhannen, has scores that fall within the PPD's interquartile range 85% of the time and within the 95% interval 99% of the time.

To better evaluate the model's consistency with particular judges, we compared each rater's pattern of rankings with the PPD. Figure 2 displays histograms of the true ratings provided by each judge. For example, we can see that Freedom House placed between 400 and 500 country-years in its lowest category, more than 600 observations in its highest category, and so on, while PACL rated over 4000 country-years as democracies and approximately 3000 cases as authoritarian regimes. Overlaying each histogram bar is a box plot of the equivalent count in the sample of draws from the fitted model's PPD. Figure 2

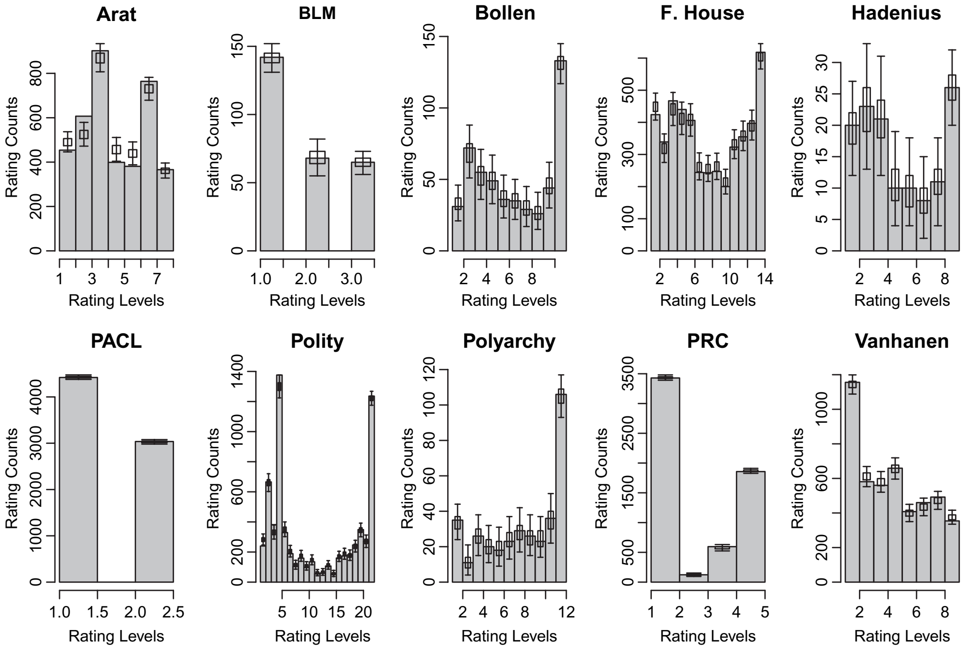

Fig. 2 Observed ratings and the posterior predictive distribution.

shows that the observed data is largely consistent with the model's DGP; most observed category-counts in every rater fall within the interquartile ranges of the posterior predictive draws and virtually every observed value falls within the 95% credible intervals of the PPD. Nevertheless, although it is not uncommon to reject the null a few times when conducting so many comparisons at once, and while the outliers we do observe lie very close to expected ranges, the fact that three of Arat's seven category-counts fall slightly outside their corresponding posterior predictive 95% credible intervals is cause for potential concern.[12] Therefore, we opted to test the sensitivity of our model estimates to the inclusion of Arat's measure in the analysis and refit the model to a data set containing every measure except Arat. Excluding Arat from the analysis has little impact on the substantive conclusions one can draw from the model. Comparing the two fitted models, we find that all but 29 of the 7558 estimated democracy scores (each z_i) have overlapping interquartile ranges, and every 95% credible interval overlaps.

Additionally, although our diagnostics provide no indication that the model DGP is at odds with the process that generated the PACL ratings, the commitment of Alvarez et al. (1996) to an either-or approach to democracy rating raises the possibility that PACL are ill suited for inclusion in the UDS. Therefore, we fit another version of the model, excluding PACL from the analysis. Again, we found that excluding a single measure has little impact on the conclusions that one can draw from the model, although removing PACL from the model does, on average, increase the confidence intervals surrounding the democracy

[12] In general, the model appears to do worse at predicting Arat's scores than it does for the other measures, regardless of the method we used to convert Arat's scores to ordinal rankings.

estimates. All but 39 of the 7558 z_i's interquartile ranges overlap across models, with 95% credible intervals overlapping in all instances.

Finally, we explored the model's sensitivity to the method we used to convert the continuous raters' scores into ordinal measures. We generated two alternate data sets by establishing arbitrary cutoffs at the deciles and vintiles of each continuous measure. The decile and vintile approaches generate sets of cutoffs that vary both in number and native-scale locations from the break points we used in the primary analysis. Again, we compared the estimates produced by the original and modified models and found little difference across specifications, with no nonoverlapping 95% intervals between the core UDS model and either alternate ordinal specification, eight differing interquartile ranges between the core and decile-based estimates, and only two nonoverlapping interquartile ranges between the primary and vintile-based approaches. Error variance estimates are also virtually indistinguishable across models. Indeed, although the interquartile ranges for Vanhanen's error estimate do not overlap across the core and vintile models, every 95% interval overlaps and every other interquartile range overlaps across all three specifications. Therefore, the UDS appear robust to the way in which we convert the continuous measures to rankings.

To summarize, the PPD distributions generated from the model largely validate the assumptions the UDS is based upon. The tests indicate that the model predicts actual rater data well and provide strong justification for including all 10 measures in the model. Moreover, the results demonstrate that any systematic bias present in the constituent measures is not so significant as to impede the model's ability to simulate the DGP driving democracy ratings. Furthermore, the model is robust to the specific method that we used to convert continuous ratings into ordinal rankings.

5 The UDS

We generated UDS for virtually all countries in the world from 1946 to 2000. Samples drawn from the estimated conditional posterior distributions of these scores and the other model parameters are available on the UDS Web site at http://www.unified-democracy-scores.org, accompanied by a tutorial that demonstrates how to use the UDS in applied analyses, taking measurement error into account.

Figure 3 shows a cross-section of UDS, for the year 2000. The dots in the figure represent mean posterior democracy scores for each country—a point estimate of the country's democracy level in 2000—and the horizontal bars depict 95% highest posterior density (HPD) regions around the estimates. An examination of the point estimates demonstrates that the measure has significant face validity. Countries' UDS tend to align closely with common perception among comparative and international relations scholars. Bastions of tyranny such as North Korea or Turkmenistan rate at the bottom of the UDS scale, whereas developed western democracies top the scale. Similarly, developing democracies like the Philippines or Honduras inhabit the middle of the scale, where many comparative scholars would place them.

A significant benefit of the UDS over their component ratings lies in their ability to estimate the measurement error of democracy. The 95% posterior density intervals surrounding each point estimate in figure 3 graphically display this estimated error. Accounting for measurement error is critically important, and historically, the issue of measurement error has been largely ignored by democracy raters. This is primarily because it is very difficult to provide estimates of measurement error without comparing multiple

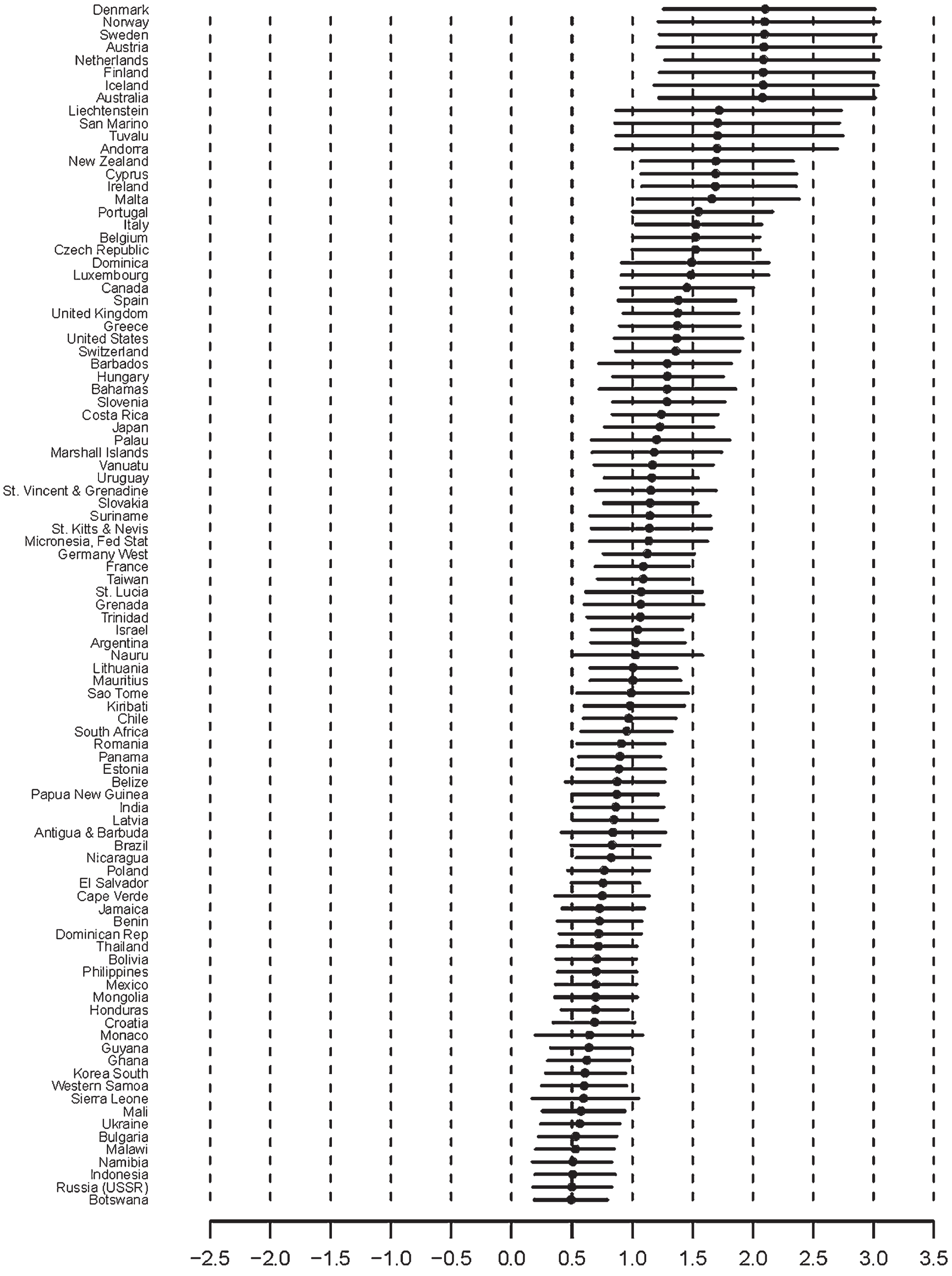

Fig. 3 *Continued*

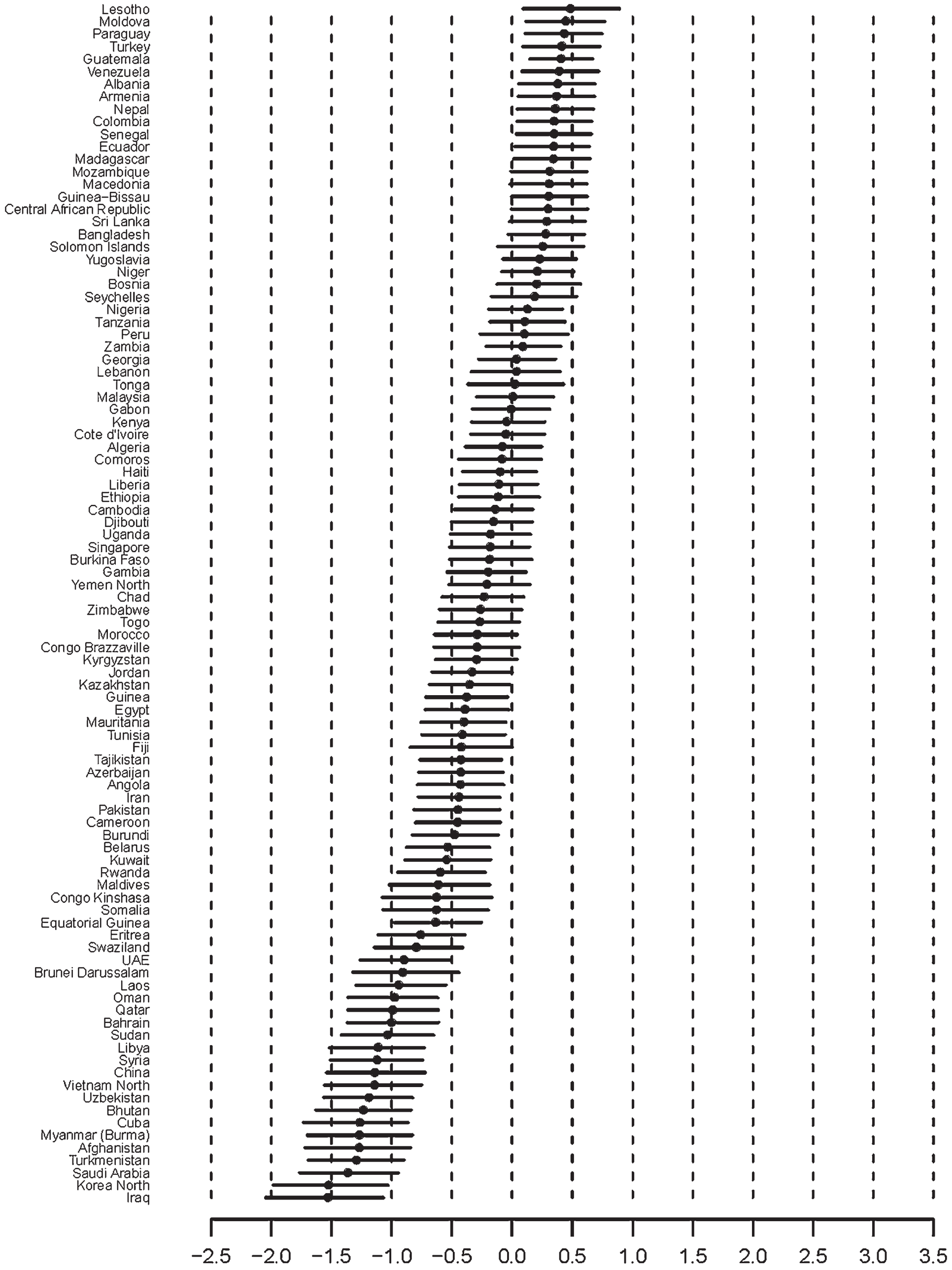

Fig. 3 Unified democracy scores for 2000.

raters' measures.[13] With the UDS, by contrast, we can estimate where we should be the most and least confident in our ability to rate democracy.

One especially striking effect of taking democracy measurement error into account is the high amount of uncertainty surrounding the UDS of the most democratic countries in the sample.[14] Error bars for developed democracies are large, indicating our uncertainty about the point estimates and the UDS' limited ability to discriminate between developed democracies. This is a result of the right truncation inherent in the individual component scales that make up the UDS. Since there is no agreed-upon method to distinguish between the level of democracy in the most democratic countries, any variance in the ratings of developed democracies among raters tends to be the consequence of idiosyncrasies (e.g., in the case of Vanhanen, a drop in voter turnout from one election to the next or, in the case of Freedom House, perhaps a poorly timed newspaper headline or rogue expert rater), not systematic changes in the level of democracy. As a whole, there is no systematic pattern in relative rankings across raters at the top end of the democracy scale. The model interprets this incoherence as measurement error and translates it into larger error bars surrounding the point estimates for developed democracies, relative to other countries.

The UDS clearly show that, in order to draw conclusions between developed democracies, the major players in the measurement of democracy need to agree on what characteristics differentiate them. Although scale truncation is a common problem in many measurement domains, the asymmetrical nature of the problem is striking in this context. Interestingly, this problem is substantially more pronounced at the top of the scale than at the bottom. The minimalist conceptualizations of democracy that our constituent raters rely upon distinguish between regimes that sport some of the baseline criteria for democracy and those that do not, but they have little to say about variations in regimes that exhibit all the commonly cited characteristics of democracy. It appears that our measures of democracy might be better described as measures of autocracy or authoritarianism. Since scholars are divided about what features make developed democracies more or less democratic than one another (e.g., parliamentary versus presidential systems, single-member versus proportional representation, etc.), the UDS is unable to distinguish between these countries. On the other hand, the minimum requirements for democracy are better articulated and more widely shared, allowing for finer distinctions at lower levels of democracy.

Figure 3 also shows that, even after accounting for measurement error, the UDS retain their ability to effectively distinguish between countries' scores. Building on multiple component scales provides researchers with a tool that balances a concern for measurement error with the ability to distinguish between countries that most reasonable people believe differ in their level of democracy. This is, perhaps, the most important contribution of the UDS. Although estimates of uncertainty are largely absent from extant scales, recent work quantifying confidence in the Polity scores provides evidence that our ability to distinguish democracy levels across countries may be troublingly low (Treier and Jackman 2008). Nonetheless, by applying our model to the patterns of agreement across raters, we can generate confidence bounds that accurately reflect the level of consensus in the field. Therefore, scholars that use the UDS—and their estimates of measurement error—in applied research will be able to make reasonably fine-grained distinctions about country's democracy levels, while holding their results up to a rigorous standard of robustness.

[13] Treier and Jackman (2008) estimated the measurement error in Polity by recombining the components of Polity with an item response model. Since component measures are rarely made publicly available for measures of democracy, their approach is not applicable for most measures.

[14] This behavior is robust to the sample year observed.

Indeed, as figure 3 makes clear, the signal to noise ratio captured by the UDS is reasonably high. For example, the unified scores show that, in the light of the verdicts of multiple raters, we should be quite confident that the United States were more democratic than such developing democracies as Brazil ($p = .96$) and India ($p = .94$) in 2000.[15] On the other hand, it is more difficult to conclude whether or not Brazil is more democratic than Honduras ($p = .73$), although Polity, Polyarchy, and Vanhanen all rank Brazil higher than Honduras. Similarly, although both Polity and Freedom House place Jordan multiple rankings above Egypt on their native scales, a concern for measurement error obscures this distinction ($p = .59$). Nonetheless, there is strong consensus among raters that both these regimes are more democratic than Syria ($p = .99$). Overall, the ability of the UDS to discriminate between cases where the weight of available evidence leads to strong conclusions and those where distinctions are more hazy demonstrates the power of incorporating as much available information as possible into one's analysis.

In fact, the relationship between the number of raters scoring a given country in a given year and the uncertainty around that country-year's estimate is evident in the model posterior. The average SD in the UDS' posterior distribution dips down in the years when there are many observations and is at its lowest in 1981 and 2000, when the number of observations is greatest, showing that our confidence in democracy score estimates does indeed increase with the available rating data. Furthermore, if we refit the model using only the three best-known raters—Freedom House, PACL, and Polity—we find that the average posterior SD around the democracy estimates jumps by 24%, indicating that adding raters generally tightens confidence intervals. Furthermore, each additional rater contributes to the precision of the UDS. We fit a series of nine-measure models, dropping each rater from the UDS in turn, and compared the posterior SDs around democracy estimates between the reduced models and the full UDS. Including Arat in the model reduces average posterior SDs by 10.5% for countries rated by Arat. The other measures' average percentage contributions to precision on cases that they rated were 10.1 (BLM), 16.6 (Bollen), 20.9 (Freedom House), 28.0 (Hadenius), 2.9 (PACL), 26.0 (Polity), 19.5 (Polyarchy), 6.9 (PRC), and 13.9 (Vanhanen). These contributions reflect a complicated combination of factors, including rater reliability and specificity, and the pattern of overlap in judges' ratings.

These findings do not imply that adding more measures substantially increases the reliability of the UDS in all cases. On average, the confidence intervals are tighter when more measures are present, but there is variation. For instance, in figure 3, although the Maldives have a similar mean democracy rating (−0.61) to Burundi (−0.47), Eritrea (−0.76), and Rwanda (−0.59), the 95% HPD region spanning the Maldives' average (0.84) is around 17% wider than the confidence regions around the point estimates for Burundi (0.72), Eritrea (0.72), and Rwanda (0.73). All four countries share three raters—Freedom House, PACL, and Polity—but the Maldives were rated by Vanhanen, whereas the other countries were evaluated by Polyarchy. The model finds Vanhanen substantially less reliable than Polyarchy (see fig. 5). Therefore, the confidence intervals around the mean UDS for the Maldives are substantially larger than those for the other countries, in 2000. This trend is consistent throughout the data; although adding reliable raters can substantially increase confidence in particular UDS, the model reacts sensibly to less reliable judges by maintaining wide confidence intervals around estimated scores.

[15] We obtain these figures by calculating $\Pr(z_i < z_{\mathrm{US}})$ for each country-year i in 2000. Using our Bayesian approach, this is simply a matter of counting the proportion of times $z_i < z_{\mathrm{US}}$ in the sample simulated from the posterior distribution of z.

6 Comparing Existing Democracy Measures

We argue that the UDS represent an improvement over isolated democracy measures because they draw upon the work of a wide variety of scholars to both improve rating accuracy and to generate estimates of score uncertainty. Therefore, it generally makes sense to use the UDS when conducting research that relies on quantitative estimates of democracy. But, for scholars that have specific theoretical reasons to use another measure, the UDS can still provide useful information to assist in their choice between democracy scores. Such scholars have few concrete empirical analyses to examine when adjudicating between measures. Fortunately, the scaling process that generates the UDS also creates compelling—and straightforward—comparative diagnostics between the various component scales that can assist in such decisions. These diagnostics allow us to compare the components of the UDS directly on the same underlying scale in terms of relative cutoffs, overall reliability, and agreement.

6.1 *Rater Cutoffs*

One tantalizing question often confronts scholars using measures of democracy: how should we interpret the scores of the various measures relative to one another? Researchers in the field often speculate what, for example, a score of three on Polity means on Freedom House's scale or where precisely PACL's distinction between democracy and autocracy falls on other popular scales. Unfortunately, current democratic measures are not synchronized with one another, making this sort of direct comparison difficult. Simple standardization techniques like the one we employed to create figure 1 are indicative of overall measure congruence but can do little to overcome this fundamental scaling issue. The multirater model employed here, by contrast, allows us to estimate the 10 raters' score cutoff points (the model's γ parameters) along a uniform scale, making direct comparisons across scores possible.

Figure 4 shows the estimated placement of the various cutoffs for each measure in relation to one another, using information from all the component measures over the entire postwar period. Each bar on the figure represents a cutoff between two score levels on the same measure. Because all raters are scaled to the UDS, we can determine where the cutoffs for these scores are in relation to one another. PACL, for example, only features a single cutoff because the scale is dichotomous. Above the cutoff, PACL rates the country a democracy, and below, it is an autocracy. Additionally, the size of the bar itself indicates the uncertainty about the score around the cutoff, as measured by the 95% HPD interval. For example, the model indicates that there is a 95% chance that PACL's democracy-autocracy cutoff falls between 0.18 and 0.25 on the unified scale. For measures with significant uncertainty with respect to ratings at any given democracy level, the bars are large, indicating that the cutoff cannot be placed reliably on the underlying UDS. Larger error bars are primarily caused by a paucity of observations. Hence, the four raters with the smallest number of observations over all—BLM, Bollen, Hadenius, and Polyarchy—clearly have the largest error bars. There is also some variance in the size of the error bars within raters caused by rater inconsistency, a paucity of observations at a particular level, or both, but as illustrated in figure 4, most of the variance is between raters.

The estimation of these cutoffs allows us to answer a number of substantively interesting questions. The first item of note is that, within raters, the error bars rarely overlap: cutoffs are typically spread out and the error bars around the cutoffs are generally tight. This means that most raters exhibit sufficient consistency for ratings on their scales to have meaning.

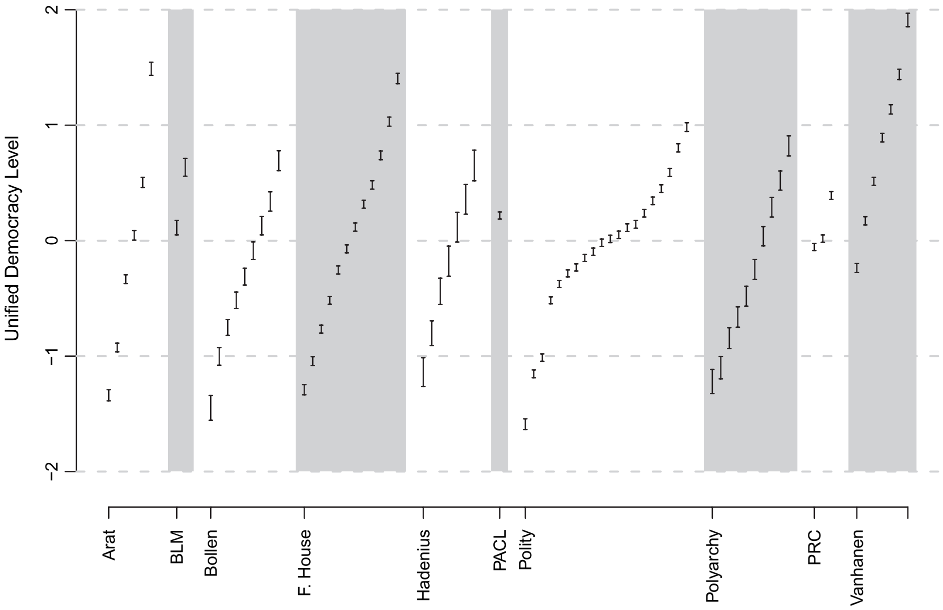

Fig. 4 Democracy measure rating cutoffs.

That is, single-level differences in ratings within a given measure are largely meaningful and only in relatively few cases do we have evidence that raters' rankings should be collapsed into fewer categories. Nonetheless, there are a number of notable discrimination problems. Take, for example, the meaning of the middle scores (from approximately −3 to 3) on the combined Polity measure. Some comparative scholars suggest that the middle of Polity is muddled and that the difference between countries that are scored one or two points apart in the middle of the Polity scale is not substantively significant and, very often, arbitrary. This complaint is well founded. The error bars around Polity scores in the middle of its scale are relatively close together, considering their size, and in fact overlap considerably. There is simply too much overlap in the true level of democracy in each of the categories in the middle of Polity for the model to distinguish cutoff locations effectively, indicating that these categories should likely be collapsed. Analyses that treat Polity as an interval scale and give identical weight to differences between scores in the middle and at the ends of the scale run the risk of drawing improper inferences. Based on this finding, one should carefully evaluate any result driven by differences in the middle of the Polity scale. Similarly, because of their relatively small sample sizes, both Hadenius' ratings and Polyarchy scores suffer from wide, overlapping, confidence intervals around their cutoffs. Users of these measures should consider collapsing categories before using these scores for inferential purposes.

We can also use figure 4 to evaluate the consistency of raters' scaling strategies. Some judges do a better job than others of setting cutoff points that smoothly span the score space. For example, Freedom House's cutoffs move across the space in a reasonably stair-step fashion, exhibiting relatively uniform distances between cutoffs across the entire scale, with some variation at the extremes. On the other hand, Polity's scores vary significantly from cutoff to cutoff. This distinct lack of uniformity in cutoff placement may reflect issues with Polity's oft-criticized score aggregation method (Gleditsch and Ward 1997; Treier and Jackman 2008). Furthermore, the estimated cutpoints of the continuous measures highlight the dangers of treating these scales as truly interval-level. Both Arat and Vanhanen's cutpoints, while evenly spaced in their native scales, lurch across the unified score space in irregular jumps, exhibiting strong evidence of non-linearity. Bollen and Hadenius demonstrate higher levels of consistency, but both exhibit some irregularity at the tails. In general, although most of our component scales demonstrate reasonably consistent cutoff placement, virtually every rater exhibits some inconsistency at the high and low ends of the democracy scale. Cutoff consistency is important because many researchers treat democracy scores as interval measures in their analyses and large nonlinearities in cutoff placement can potentially bias results.

The score cutoff estimates also allow us to examine the validity of some commonly used democracy measurement ''rules of thumb'' in comparative politics and international relations. Take, for example, the tendency of researchers to dichotomize Polity to generate strict democracy-autocracy scores from the ordinal measures. A number of theories rely on the presence of democracy, not on level of democracy, to explain phenomena, requiring such an approach.[16] A cutoff used by a variety of international relations scholars is to code nations with democracy scores greater than 6 or 7 on the combined Polity scale as democracies and all other countries as non-democracies (Ray 2000), whereas the PACL raters explicitly conceptualize democracy as an either-or concept and rate all countries in a binary fashion. We can evaluate the consistency of these two approaches to democracy dichotomization using the cutoffs in figure 4. The figure shows that the use of 6 on Polity to dichotomize democracy is reasonably consistent with the PACL definition of democracy: the PACL cutoff lies somewhere near 5 on the polity scale.

6.2 *Rater Reliability*

Comparing rater cutoffs helps us examine relative meaning across democracy measures, but we may be even more interested in judging raters' relative levels of reliability. Indeed, much discussion in the literature regarding the strengths and shortcomings of various measures touches on the question of overall reliability.[17] The multirater ordinal probit generates estimates of each rater's tendency to make idiosyncratic mistakes, parameterized as each rater's error variance σ_j^2. These estimates capture rater reliability (reliability is simply the inverse of the error variance) and are a function of the level of agreement between raters across the country-year sample. Figure 5 plots estimated rater-specific error variances for all judges with 95% HPD intervals; high variance means more errors and less reliability and vice versa.

The overall picture from the reliability comparison is encouraging for those scholars who argue for democracy scale agnosticism (Adcock and Collier 2001). With the exception of Arat and Vanhanen, which exhibit substantially higher error variances than the remaining measures, all the scores demonstrate similar levels of reliability.[18] The big three raters—Freedom House, PACL, and Polity—all demonstrate moderate to high reliability. And, although we can say with some confidence that PACL is more reliable than Polity, and Polity more reliable than Freedom House, those differences are minor, and all three raters exhibit error variances that are quite small in respect to the range of the UDS scale. PACL,

[16] Findings of the democratic peace literature spring to mind.

[17] For example, Bollen (1980) places measure reliability at the center of his analysis.

[18] Robustness checks indicate that our decision to discard continuous rating information is not responsible for the high estimated error variances for Arat and Vanhanen. Instead, their reliability is likely the result of their chosen measurement strategies. We discuss this issue in depth in the Web appendix.

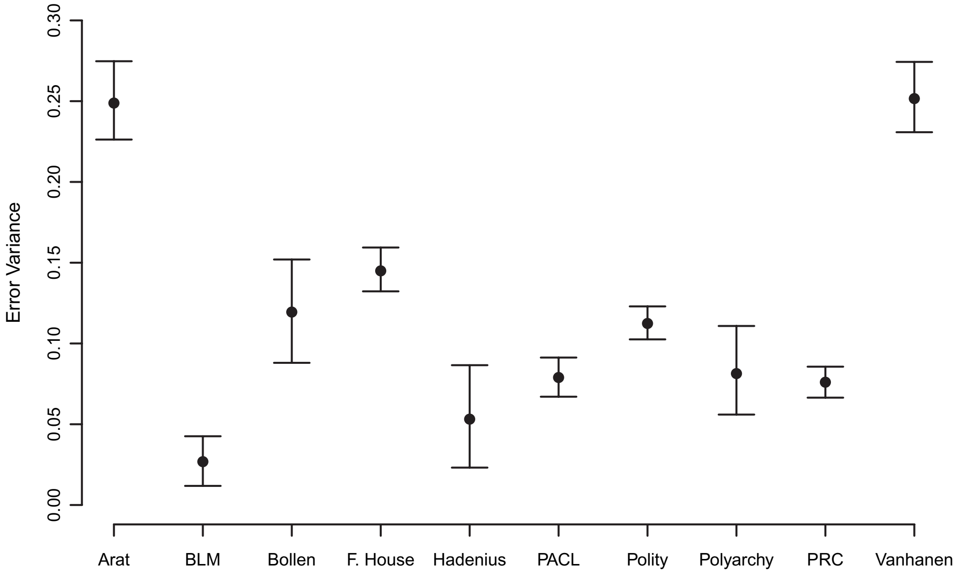

Fig. 5 Democracy measure error variance.

furthermore, is one of the most reliable raters in the UDS, making relatively few mistakes when categorizing countries and is statistically more reliable than either of the other two popular measures. Although PACL is explicitly not continuous in construction, it behaves in a manner consistent with a conceptualization of democracy as a continuous latent variable. Researchers choosing between the major players in democracy measurement have little reason to make distinctions based on reliability and, at least on this criterion, should be well served by any of the three most commonly used raters. Similarly, PRC provides a nice mix of wide case coverage and low estimated error variance.

However, being a large, multi-decade rater with global coverage does not automatically guarantee reliability. For example, Vanhanen's measure, which provides extensive coverage, generates ratings that are often inconsistent with the evaluations of the other judges, resulting in a high error variance estimate. Furthermore, some of the most reliable raters in the sample are smaller, more focused, projects like Polyarchy by Coppedge and Reinicke (1991), Hadenius' (1992) scale, and especially measure of democracy in Central America by Bowman, Lehoucq, and Mahoney (2005). These projects all feature limited country coverage, minimal time spans, or both. The high reliability of these small scales may be surprising to readers, especially those who believe that differences between democracy measures are largely a result of validity issues and systematic bias, rather than reliability problems and random errors as our model assumes. In fact, if focused measurement projects tended to outperform the large-*N* measures primarily in terms of validity, our model would likely find the area experts' scores unreliable. Therefore, the model's tendency to estimate low error variances for targeted raters provides at least some evidence that a reliability-based approach to modeling democracy rating is appropriate.

Of course, to many comparative scholars, the reliability of these small-*N* projects will not be surprising at all. Groups focusing their research on certain periods or regions are likely to know their areas well and may be able to devote greater resources to each individual score than an extended-coverage measurement project. The BLM measure, the most reliable component of the UDS, provides a case in point. It was intentionally created in response to perceived data-induced measurement error, ''grow[ing] out of the use of inaccurate, partial, or misleading secondary sources'' (Bowman, Lehoucq, and Mahoney 2005, 940). The authors, Central America experts, went to great lengths to use only primary sources when scoring five Central American countries for 100 years.

The reliability of small projects like BLM raises intriguing possibilities when combined with the UDS system. Previously, because most scholars employing quantitative democracy scores are engaged in large-*N* statistical analyses requiring wide case and time coverage, there existed little incentive to produce small-scale quantitative democracy scales. Score aggregation methods like the UDS approach demonstrate the potential a network of dedicated case-focused scholars have to improve reliability in democracy measurement. Using the existing large-scale projects as a sort of measurement glue, one can incorporate the work of numerous, highly reliable, small-*N* comparative scholars to substantially reduce the reliability issues of democracy scores.

Although case coverage appears to have an important effect on rater reliability, few other major patterns emerge from the error variance estimates. For example, there is no relationship between the number of indicators the judges used to construct their measures—as reported in Table 1—and their estimated error variances ($r = -.23$).[19] Similarly, while raters that Munck and Verkuilen (2002) cite as having ''clear and detailed coding rules,'' including Hadenius, PACL, and Polity, have some of the lowest estimated error variances of the included measures, Vanhanen, which Munck and Verkuilen (2002) also describe in this way, is likely to be the least reliable rater.

Although the results here may quell some fears among scholars about the reliability of common measures, it is still necessary to reintroduce some caution. From a reliability perspective, it is best to use the information from all the scales present in the UDS rather than individual measures. Measurement error and mistakes are still a serious problem for individual measures, even among the more reliable scales. As we previously demonstrated, even reliable raters make mistakes for certain country-years. The UDS is less likely than the individual measures to be misled by such mistakes. Given the model assumptions, the UDS is known to be at least as reliable as the most reliable of the component measures and in almost all cases is significantly more so.

6.3 *Rater Differences*

Throughout this paper, we have argued that, despite the high correlations between the 10 existing measures of democracy, there are discrepancies between them that matter to the applied researcher. Indeed, as we previously noted, this fact has been demonstrated in the literature (Elkins 2000; Casper and Tufis 2002). So how common are discrepancies between these measures? In the three most commonly used measures of democracy, Freedom House, PACL, and Polity, we find discrepancies are a common occurrence. Using model-generated estimates of rater perceptions (the t parameters), we can determine when raters provide statistically distinguishable ratings of country-years' democracy levels.[20] Based on data

[19] The components we refer to are the ones listed in Table 1. However, Munck and Verkuilen (2002) list subcomponents for a subset of these measures: Arat (8), BLM (10), Bollen (6), Freedom House (22), and Hadenius (6). If we treat these subcomponents as the ''components'' of these scores, we find even less of a relationship between estimated error variance and component number ($r = .026$).

[20] The online appendix explains how we statistically classified rater discrepancies.

from our model, of the 3929 country-years when both Freedom House and Polity provide scores, 2424, or around 62%, of the scores are statistically distinguishable from each other. Of the 4,698 country-years for which both Freedom House and PACL provide scores, 556, or about 12%, of the scores are statistically distinguishable from each other. Furthermore, of the 6481 country-years, when both PACL and Polity provide scores, 655, about 10% of the scores are statistically distinguishable. Finally, although there can be significant discrepancies across pairs of raters, cases with differences generally vary across the raters compared. Looking at the 3929 country-years when all three measures are present, only 136, or less than four per cent, of the scores are different across all three measures.[21]

It is important to put these findings in context; the majority of discrepancies, although statistically significant, are substantively small. Nonetheless, certain differences may have important substantive implications. For instance, using 0.22, the estimated posterior mean of PACL's single cutpoint, as our cutoff, we calculated the number of cases for which two raters have a 95% or better chance of providing differing opinions of a country-year's democratic status. We found that Freedom House and Polity disagreed on around 4% of the cases they both rated, whereas Freedom House and PACL disagreed over 6% of the time, and PACL and Polity differed on almost 7 of cases. Furthermore, there are cases where particular raters provide estimates that are substantially at odds with the judgments of all the other scholars. We argue that these discrepancies are evidence of potentially serious problems in research using any one of these measures of democracy. The UDS provide the information necessary to identify these cases: if a rater's perception of a country-year's democracy scores is statistically distinguishable from the UDS then we know that the rating represents an unusual observation. Researchers intent on using an isolated democracy score in their research should use the UDS to identify such observations in their data sets and evaluate the robustness of their results to these observations, just as they would to any other outlier. These outlying scores are not uncommon: according to the model, more than 21% of Freedom House's 4703 ratings have a 95% chance of being strictly greater than or strictly less than the UDS composite score, whereas almost 3% of PACL's 7457 scores meet this criterion,[22] and almost 14% of Polity's 6577 scores are statistically distinguishable from the corresponding unified estimate.

7 Conclusion

Comparative and international relations scholars no longer need to make arbitrary decisions about the democracy measure that they include in their quantitative analyses. Instead, the techniques introduced here allow scholars to combine the work of many democracy raters into a single set of scores. This approach may be generalized to other domains where multiple, yet complementary, measures exist, such as political sophistication or state economic openness. The UDS also reemphasize the importance of incorporating estimates of error into measures of unobservable concepts (Treier and Jackman 2008). Even using the cumulative knowledge of all the judges discussed here, measurement error can still be a barrier to differentiation between democracies, and the problem is even more profound in individual measures. The UDS' framework provides an ideal way to reduce measurement

[21] In contrast to rater error estimates, which are not terribly sensitive to the number of rater categories, these figures do tend to grow with the number of rater cutoffs, all else equal. Thus, the higher congruence when PACL is involved is, to some extent, an artifact of the scale's limited specificity.

[22] These figures do tend to grow with the number of rater cutoffs, all else equal.

error in empirical work on democracy; more information, ultimately, is the only real solution for uncertain measures.

The UDS provide a jumping-off point for a number of related research agendas. First, to improve model efficiency, it would be useful to develop a semiparametric hybrid model where we relax the assumption that t_{ij} and y_{ij} are linearly related without throwing out continuous information prior to estimation, allowing us to work with continuous data that do not meet the basic requirements of interval measurement. This new approach would allow us to fit a model that is more efficient than the purely ordinal model that we use here without making overly strong assumptions about continuous measures. Second, it is possible to incorporate covariates in the multirater ordinal probit and to treat functions of these covariates as additional raters. We are currently investigating the practicality of using purely objective institutional measures, such as constitutional features, to generate reliable democracy scores that are consistent with existing measures. A function of objective measures capable of mimicking existing raters would have the potential both to reduce the costs associated with democracy measurement and to help us unpack what remains a highly subjective, often impenetrable, process. Third, the exceptional reliability of small-scale measurement projects, like BLM's contribution, highlights the potential that area scholars have to improve the quantitative measurement of democracy. As the UDS evolve, the inclusion of more such measures would provide substantial reductions in our uncertainty around estimates. Finally, the multirater approach described here could be expanded to take systematic bias into account, improving on the random-error model used to create the UDS. Although previous research has broached this topic (Bollen and Paxton 2000), there has been no effort to produce a set of synthesized democracy scores directly from such a model. These bias-corrected UDS would provide a tool for students of democracy that would be both more reliable and more valid than currently available measures.

References

Adcock, Robert, and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95:529–46.

Alvarez, Michael, José Antonio Cheibub, Fernando Limongi, and Adam Pzeworski. 1996. Classifying political regimes. *Studies in Comparative Political Development* 31:1–37.

Arat, Zehra F. 1991. *Democracy and human rights in developing countries*. Boulder, CO: Lynne Rienner.

Bollen, Kenneth A. 1980. Issues in the comparative measurement of political democracy. *American Sociological Review* 45:370–90.

———. 2001. *Cross-national indicators of liberal democracy, 1950–1990*. 2nd ICPSR version. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/02532.xml.

Bollen, Kenneth A., and Robert W. Jackman. 1989. Democracy, stability, and dichotomies. *American Sociological Review* 54:612–21.

Bollen, Kenneth A., and Pamela Paxton. 2000. Subjective measures of liberal democracy. *Comparative Political Studies* 33:58–86.

Bowman, Kirk, Fabrice Lehoucq, and James Mahoney. 2005. Measuring political democracy: case expertise, data adequacy, and Central America. *Comparative Political Studies* 38:939–70.

Casper, Gretchen, and Claudiu Tufis. 2002. Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated datasets. *Political Analysis* 11:1–11.

Cheibub, José, Jennifer Gandhi, and James Vreeland. 2010. Democracy and dictatorship revisited. *Public Choice* 143:65–101.

Clinton, Joshua D., and John S. Lapinski. 2006. Measuring legislative accomplishment, 1877–1994. *American Journal of Political Science* 50:232–49.

Clinton, Joshua D., and David E. Lewis. 2007. Export opinion, agency characteristics, and agency preferences. *Political Analysis* 15:3–20.

Collier, David, and Robert Adcock. 1999. Democracy and dichotomies: A pragmatic approach to choices about concepts. *Annual Review of Political Science* 2:537–65.

Coppedge, Michael. 2002. Democracy and dimensions: Comments on Munck and Verkuilen. *Comparative Political Studies* 35:35–9.

Coppedge, Michael, Angel Alvarez, and Claudia Maldonado. 2008. Two persistent dimensions of democracy: Contestation and inclusiveness. *The Journal of Politics* 70:632–47.

Coppedge, Michael, and Wolfgang H. Reinicke. 1991. Measuring polyarchy. In *On measuring democracy: Its consequences and concomitants*, ed. Alex Inkeles, 47–68. New Brunswick, NJ: Transaction.

Dahl, Robert A. 1972. *Polyarchy: Participation and opposition*. New Haven, CT: Yale University Press.

Elkins, Zachary. 2000. Gradations of democracy? Empirical tests of alternative conceptualizations. *American Journal of Political Science* 44:287–94.

Freedom House. 2007. *Freedom in the World*. http://www.freedomhouse.org.

Gasiorowski, Mark J. 1996. An overview of the political regime change data set. *Comparative Political Studies* 29:469–83.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian data analysis*. 2nd ed. Boca Raton, FL: Chapman & Hall.

Gleditsch, Kristian S., and Michael D. Ward. 1997. Double take: A reexamination of democracy and autocracy in modern polities. *The Journal of Conflice Resolution* 41:361–8.

Hadenius, Axel. 1992. *Democracy and development*. Cambridge: Cambridge University Press.

Jackman, Simon. 2004. What do we learn from Graduate Admissions Committees?: A multiple-rater, latent variable model with incomplete discrete and continuous indicators. *Political Analysis* 12:400–24.

Johnson, Valen E. 1996. On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association* 91:42–51.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal data modeling*. New York: Springer.

Marshall, Monty G., Keith Jaggers, and Ted Robert Gurr. 2006. *Polity IV: Political regime characteristics and transitions, 1800–2004*. http://www.cidcm.umd.edu/polity/.

Munck, Gerardo L., and Jay Verkuilen. 2002. Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative Political Studies* 35:5–34.

Przeworski, Adam, Michael Alvarez, José Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and development: Political regimes and economic well-being in the World, 1950–1990*. Cambridge: Cambridge University Press.

Quinn, Kevin M. 2004. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis* 12:338–53.

Ray, James Lee 2000. Democracy: On the level(s), does democracy correlate with peace? In *What do we know about war?*, ed. John Vasquez, 299–316. Lanham, MD: Rowman and Littlefield.

Reich, Gary. 2002. Categorizing political regimes: New data for old problems. *Democratization* 9:1–24.

Treier, Shawn, and Simon Jackman. 2008. Democracy as a latent variable. *American Journal of Political Science* 52:201–17.

Vanhanen, Tatu. 2003. *Democratization: A comparative analysis of 170 countries*. New York: Routledge.