

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

Daniel Pemstein[†] Stephen Meserve[‡] James Melton[‡]

February 7, 2008

Abstract

Using a Bayesian latent variable approach, we synthesize a new measure of democracy, the Unified Democracy Scores (UDS), from ten extant scales. We accompany this new scale with quantitative estimates of uncertainty, provide estimates of the relative reliability of the constituent indicators, and quantify what the ordinal levels of each of the existing measures mean in relationship to one another. Our method 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 authors are listed in reverse alphabetical order, indicating equal contribution to the article. The authors would like to thank Bill Bernhard, Zach Elkins, Brian Gaines, Jim Kuklinski, Kevin Quinn, seminar participants at the University of Illinois, and participants at the 24th Annual Meeting of the Society for Political Methodology for helpful comments.

[†]Doctoral Fellow, The Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 and Ph.D Candidate, University of Illinois at Urbana-Champaign.

[‡]Ph.D. Candidate, University of Illinois at Urbana-Champaign, 361 Lincoln Hall, 702 South Wright Street, Urbana, IL 61801

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 & 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 & Collier 2001). This is somewhat surprising considering the wide array of conceptualization, measurement, and aggregation techniques employed by the creators of these scales (Munck & Verkuilen 2002). However, despite the similarities in these scales, the subtle differences between them can and do affect substantive results (Casper and Tufis 2002, Elkins 2000). Moreover, while 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 & Jackman 2007). Even dismissing these two problems, the pragmatic recommendation to select the single measure that best operationalizes democracy for the question at hand (Collier & 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 uncertainty 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 inter-coder validation across major democracy scales. The UDS are also flexible and incorporate information both from measures spanning a handful of country-years and multi-year 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 confidence intervals around the UDS reflect these disagreements and allow researchers to deal with these cases in a reasoned and systematic manner. In addition, modeling the rating process has merit beyond the creation of the UDS themselves. Our model makes possible the direct analysis of various scales in relation to one another, helping researchers to calibrate cut-points 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.

In what follows, we begin by describing the building blocks of the UDS, starting with a discussion of the ten existing democracy scales that form the foundation of the unified scores. We then develop and formalize a model of the democracy rating process and link this conceptual framework to a statistical measurement model that effectively synthesizes

the contributions of diverse democracy raters into a single unified set of scores. We explain how to identify and estimate this model, fit the model to the available data, summarize the resulting UDS and demonstrate that the UDS improves inference by reducing and quantifying uncertainty about measurement. We then use the wealth of information generated during the estimation of the UDS to perform heretofore impossible comparisons between the ten component measures. Specifically, we place the democracy scores provided by each rater on a single latent scale, generating diagnostic information useful to researchers interested in working directly with the base measures, compare levels of reliability across the various ratings, and investigate the prevalence of divergent estimates of democracy within the base measures.

1 A Plethora of Measures

The UDS incorporate information from ten measures of democracy: Arat (1991), Bowman, Lehoucq & Mahoney (2005) (BLM), Bollen (2001), Freedom House (2007), Hadenius (1992), Przeworski, Alvarez, Cheibub & Limongi (2000) (PACL),¹ Marshall, Jaggers & Gurr’s (2006) Polity scores, Coppedge & Reinicke’s (1991) Polyarchy scale, Gasiorowski’s (1996) Political Regime Change measure (PRC), and Vanhanen (2003). All ten measures are based on similar underlying conceptualizations of democracy. Munck & Verkuilen (2002, 9), discussing nine out of the ten 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 to maintain this wide scope. Other measures, like BLM and PRC,

¹The original scores provided by Przeworski et al. (2000) include 141 countries from 1950 to 1990. We use an extended version of this dataset that covers the years between 1950 and 2000 for as many as 189 countries (Cheibub & Gandhi 2004)

provide limited coverage but are based on in depth analyses of primary source material. Most of the scales rely, at least to some degree, on subjective expert ratings but Vanhanen (2003) takes a completely objective approach, generating an index from readily observable indicators. The vast majority of measures use some form of additive system to incorporate individual indicators into a final democracy score but certain raters, specifically Bollen (2001) and Coppedge & Reinicke (1991), use more sophisticated techniques to help deal with issues of aggregation and reliability. Finally, the raters vary in the underlying characteristics they choose to incorporate in their final scores. Famously, for example, Freedom House is known to include an evaluation of civil rights in its measurement of democracy, while Polity does not. Thus, each judge taps Dahl's (1972) conceptualization of democracy differently and provides potentially valuable information not available in other scores. Table 1 summarizes the ten constituent measures and describes each rater's country-year coverage,² scale range, and constituent components.

[Table 1 about here.]

[Figure 1 about here.]

Notwithstanding their differences, does it really matter which measure scholars use in their research? Both large-N studies (Casper & Tufis 2002, Elkins 2000) and case evidence suggest that it can. Figure 1, which displays standardized³ Freedom House, PACL, and Polity scores for Spain, Russia, Fiji, and Burundi, can help us explore this question in more detail. In general, the available democracy measures correlate highly, as shown in table 2. This fact is often used as evidence of the (convergent) validity of the measures (Bollen 1980). While there is certainly disagreement on this point (Munck & Verkuilen 2002, 29), the correlations between measures are, at the very least, evidence that the raters are measuring roughly the same concept (Adcock & Collier 2001). The example of Spain in

²Some raters provide pre-1946 scores, but we restrict our analysis to post World War II.

³We standardize scales in figure 1 by normalizing each rater's score to the (0, 1) interval. While it serves our purposes here, this is a crude approach and the model introduced in section 2 provides a more logical basis for score standardization.

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. Although there are minor differences in the standardized scores, the general impressions given by all three measures about the level of democracy in Spain are the same. This is an excellent example of the overall face validity demonstrated by these measures. Spain is not an isolated case and reflects the convergence in democracy measures across the majority of the country-years in the dataset. For most countries in most years, like Spain, the various measures appear to measure the same underlying concept.

[Table 2 about here.]

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, the Russian legislature. As a result of these two events, Freedom House lowered its rating of Russia's level of democracy, 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.⁴ Russia seems to be a case where two measures looked at the same information and came to two different conclusions. Freedom House thought the events from 1996 on indicated a weakening of democracy, while Polity thought these same events led to a strengthening of

⁴It is hard to tell what caused Russia's reported level of democracy to decrease because we lack access to the component scores for Freedom House before 2006.

democracy. The varying measurement strategies employed by Freedom House and Polity both rely on information giving insight into the level of democracy in Russia. The raters make judgments that are sensible, but incomplete. Both measures cannot be correct but neither have both judges completely missed the mark. 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, while Freedom House and Polity both rank it as democratic, especially during the period before the 1987 coup. PACL ranks Fiji as an authoritarian regime based on their type II error rule. In other words, since the Fijian Alliance won every election until 1987,⁵ 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 1970–1987 than as a democratic regime. On the other hand, Polity and Freedom House both look at the Fijian government from 1970–1987 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. In this scenario, raters simply do not have enough information to classify the Fijian regime from 1970–1987, because they could not observe the counterfactual where the Indo-Fijians won an election during the time period in question. Again, we are not in a position to arbitrate definitively between measures and a more honest assessment of democracy level in this case 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, and it seems most likely that none of these measures is accurate. PACL scores the Burundian regime during this time as democratic, Polity scores it as transitional, between authoritarian and democratic, and

⁵ 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.

Freedom House scores it as authoritarian. In some sense, each of these raters is correct. The President of Burundi, a Hutu, was assassinated in 1993 by Tutsi army officers, leading to civil war, but the Hutus remained in power after another Hutu president was elected in 1994. The Hutu regime remained in power until 1996, when the Tutsis regained control through a successful coup. The elections and Hutu succession of office in 1994 seems to support PACL, but the instability in government and continual fight for power by both the Hutus and the Tutsis seems to 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 supports 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. 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 looking at raw scores. However, a number of countries demonstrate each of the patterns analyzed in figure 1, though the most common pattern is that of general agreement with minor discrepancies represented by Spain. As we shall demonstrate in our analysis, we can statistically distinguish between at least two raters' scores in as many as 50 per cent 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. To address this, we outline a model that incorporates information from all scales into a single set of scores, reduces sensitivity to raters' random errors, and reflects the field's level of consensus and confidence in democracy ratings.

2 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 & Adcock 1999). Yet, of the ten measures we consider here, all but one provide ordinal or continuous estimates. Thus, while the field is divided on this topic, it makes sense to bend to the will of the majority when evaluating existing indices. Similarly, while scholars have argued that democracy is a multi-dimensional entity (Munck & Verkuilen 2002, Coppedge 2002), the raters we examine here all provide a single summary value for each observation. And, while many of these judges also publish scores for various dimensions or sub-components of democracy, democracy measures are generally used as simple more-or-less ratings in applied work. Therefore, following Bollen & Jackman (1989) and Treier & Jackman (2007), we model each indicator as an approximation to an unobserved— or latent—continuous unidimensional variable.⁶

Specifically, each of the $j = 1, \dots, m$ judges provides a rating t_{ij} of the level of democracy in each of $i = 1, \dots, n$ 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

$$t_{ij} = z_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma_j^2) \tag{1}$$

⁶The dichotomous measure of Alvarez, Cheibub, Limongi & Pzeworski (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 Alvarez et al. (1996) measure 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.

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 non-systematic is clearly an oversimplification, but this assumption provides a useful starting point for modeling the measurement process. While complemented by other research (see, e.g. Bollen & 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 that is largely consistent with normally distributed random error. Similarly, as Bowman, Lehoucq & 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, while it is often possible to identify the particular "bias" guiding a rater's judgement in a given case, equation 1 nonetheless represents a reasonable, if imperfect, model of the overall pattern of ratings.

In addition, in so far as the true data generating process 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 . 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 large variance σ_0^2 . Together, equation 1 and this prior imply that z_i has a normal posterior distribution with mean

$$\frac{\sum_{j=1}^m \frac{t_j}{\sigma_j^2}}{\frac{1}{\sigma_0^2} + \sum_{j=1}^m \frac{1}{\sigma_j^2}} \quad (2)$$

and variance

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

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. 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 of the available information and fully capitalizing on the efforts of other researchers can reduce the uncertainty around our democracy estimates. After accounting for measurement error, we may be unable to use existing scales to distinguish between all but the most democratically disparate polities (Treier & Jackman 2007). By combining information we can substantially improve on this state of affairs and rate democratic levels with increased confidence.

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. We use a technique, multi-rater ordinal probit, originally developed to compare the performance of multiple essay graders (Johnson 1996, Johnson & Albert 1999).⁷ The indicators we analyze are ordinal and each judge provides a rating placing a subset of the n country-years into one of K_j ordered categories.⁸ The scales do not

⁷The notation that follows borrows liberally from Johnson & Albert (1999).

⁸ The scores reported by Arat (1991), Bollen (1980), Hadenius (1992), and Vanhanen (2003) are continuous. We discretize these scores to allow for a direct comparison with the remaining measures, by establishing arbitrary cutoffs at the deciles of each measure. These cutoffs were chosen to maximize the number of observations at each level of the discretized measures, making estimation easier. A mixed estimation approach such as that described by Quinn (2004) would avoid the need to discard information in the continuous measures but we leave this task to future research.

all need to use the same number of categories and, indeed, the measures we examine vary substantially in category count, ranging between two 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 datasets 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.⁹ We assume that judges imperfectly perceive the true level of democracy in country-year i , as described in equation 1, when producing their ordinal scores. We call the array of judgments $\mathbf{y} = \{y_{ij}\}$ where y_{ij} describes judge j 's ranking of country-year i and assume that judge j places country-year i in category c if $\gamma_{j,c-1} < t_{ij} \leq \gamma_{j,c}$ where $\gamma_{j,c-1}$ and $\gamma_{j,c}$ are judge-specific ranking cutoff points. These assumptions allow us to write the likelihood function for the rating matrix y as

$$L(\mathbf{z}, \gamma, \{\sigma_j^2\}) = \prod_{i=1}^n \prod_{j \in C_i} \left[\Phi \left(\frac{\gamma_{j,y_{ij}} - z_i}{\sigma_j} \right) - \Phi \left(\frac{\gamma_{j,y_{ij}-1} - z_i}{\sigma_j} \right) \right] \quad (4)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Following Johnson & Albert (1999), we identify the model using a Bayesian estimation approach, and adopt proper prior distributions for \mathbf{z} and each σ_j^2 . Specifically, we assume standard normal prior distributions for each latent trait z_i and inverse-gamma prior densities for the rater variance parameters.¹⁰ We estimate the model using the Markov chain Monte

⁹A logical, but somewhat unconventional side-effect of the ability to work with scales of varying breadth is that confidence in the final democracy estimates can vary substantially across country-years. Thus, this approach allows scholars to work with measures that reflect the current state of research in the field: estimates of democracy for country-years rated by numerous judges or by high-reliability raters are deemed more trustworthy than those scored by only two or three measurement teams, or less reliable raters.

¹⁰We set the parameters of the inverse-gamma priors to $\lambda = 0.2$ and $\alpha = 0.1$, which indicates an ex ante expectation that all of the variance parameters fall on the interval (0.01, 4.0). Our results are robust to perturbations of these parameter values. The standard normal priors on the latent democracy score vector \mathbf{z} set the scale of the parameter estimates, but altering the variance parameters of these priors has no impact on substantive conclusions.

Carlo (MCMC) algorithm described in Johnson & Albert (1999, pp. 166). We ran the algorithm for one 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 joint posterior distribution of the model parameters. Standard MCMC diagnostics for the sample are consistent with Markov chain convergence.

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 in equation 2, these estimates average across individual judges’ contributions, weighting each score by its error variance. Furthermore, we can provide posterior 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. Thus, we can extend the analysis of measurement error that Treier & Jackman (2007) perform on the Polity IV scores to an aggregation of a number of extant measures of democracy. In addition, this model lets us estimate the posterior distributions of the judge variance parameters ($\{\sigma_j^2\}$). These estimates describe the relative precision of the various democracy scales and provide a way to quantify the consistency of the differing judgments employed in the aggregate measure. 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 multi-rater 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.’s (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.

3 The UDS Measure of Democracy

We generated UDS for virtually all countries in the world from 1946–2000. Samples drawn from the estimated posterior distributions of these scores and the other model parameters are available on the UDS website at <http://www.clinecenter.uiuc.edu/research/affiliatedresearch/UDS/>.

[Figure 2 about here.]

Figure 2 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 per cent credible intervals (highest posterior density 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 democracy scale while developed western democracies top the scale. Similarly, developing democracies like Mexico or India inhabit the middle of the scale, where most comparative scholars would place them. This is to be expected because UDS are a synthesis of information from all measures. It is nonetheless important to point out that UDS visually conform to conventional expectations.

A significant benefit of the UDS over their component ratings lies in their ability to estimate the measurement error of democracy. The 95 per cent posterior density intervals surrounding each point estimate in figure 2 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, primarily because it is very difficult to provide estimates of measurement error without comparing multiple raters’ measures. 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.¹¹ Error bars for developed democracies are large, indicating our 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. Specifically, nearly all the component scales rate developed democracies with their highest rating for all or almost all years. There is no agreed-upon method to distinguish between level of democracy among the most democratic countries. These problems reflect issues with the underlying scales, however, and are merely inherited by the UDS. 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 those countries at the top. While 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 of the commonly cited characteristics of democracy. It appears that our measures of democracy might be better described as measures of autocracy or authoritarianism.

Figure 2 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

¹¹This behavior is robust to the sample year observed.

democracy levels across countries is troublingly low when we are forced to rely on a single measure. Indeed, the 95 per cent confidence intervals around the Polity-derived democracy scores reported by Treier & Jackman (2007, 12) are so large that those of established Western democracies overlap with those of semi-democracies like Iran and Singapore. In contrast, as figure 2 makes clear, the UDS provide much tighter intervals around estimates and allow researchers to deploy democracy measures in their research with significantly improved confidence over what one can do with a single measure.

Furthermore, Treier & Jackman (2007, 13) show that, at conventional 95 per cent levels of statistical significance, the Polity scores cannot say that the democracy levels of Russia and Indonesia were lower than that of the United States in 2000. In fact, 70 of the 153 countries included in their analysis—or 46 per cent—were statistically indistinguishable from the United States. The UDS are able to improve on this state of affairs and only 60 of the 191 countries in our analysis for 2000, or 31 per cent, are statistically indistinguishable from the United States at the 95 per cent level.¹² One should be careful to distinguish these statements from the confidence intervals displayed in figure 2. Because of the covariance across scores, we are often able to statistically distinguish between countries with overlapping confidence intervals. Indeed, the unified scores indicate that, in the light of the verdicts of multiple raters, we should be confident that the United States were more democratic than such countries as Brazil and India in that year.¹³ The enhanced discrimination ability of the UDS demonstrates the advantage of capitalizing on the significant efforts of other scholars and incorporating as much available information as possible into one’s analysis.

[Figure 3 about here.]

¹²This improvement is even more impressive when one considers that the most of the additional 38 countries included in the UDS are highly democratic island and other small states.

¹³We obtain these figures by calculating $\Pr(z_i > z_{US})$ for each country-year i in 2000. Using our Bayesian approach, this is simply a matter of counting the proportion of times $z_i > z_{US}$ in the sample simulated from the posterior distribution of \mathbf{z} . Thus, for example, our analysis tells us that, according to our raters, the probability that India is more democratic than the United States is 3.6 per cent, while the probability that India is more democratic than Iran or Singapore is essentially 100 per cent.

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 plainly evident in the model posterior. Figure 3 graphs component measure availability over time and compares changing sample size to the average standard deviation in democracy score estimates in each year. The average standard deviation in the UDS' posterior distribution clearly 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. This does not imply that adding more measures substantially increases the reliability of the UDS or tightens the confidence intervals in all cases. On average, the confidence intervals are tighter when more measures are present, but there is variation. For instance, in figure 2, Comoros, Congo, and Peru all have confidence intervals about 0.7 units wide, but Iceland, and Malta's confidence intervals are noticeably larger at closer to 2.0. All five of these countries are rated by four of our measures, but the former three are rated by Freedom House, PACL, Polity, and Polyarchy while the latter two are rated by Vanhanen instead of Polity. The model finds Vanhanen substantially less reliable than Polity (see figure 5 in section 4.2). Therefore, the confidence intervals around the UDS for the latter two countries are much larger than the confidence intervals for the former three. This trend is consistent throughout the data; while 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.

4 Comparing Existing Democracy Measures

We argue that the UDS represent an improvement over isolated democracy measures because they build on existing scholarship, drawing 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 current measures of democracy. Scholars who struggle to make distinctions between measures of democracy have few concrete empirical analyses to examine when adjudicating between measures. In general it appears that the choice of measure is often one of convenience or subfield norms. At best, this decision requires careful evaluation of the existing indices on a number of largely subjective criteria. 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 component measures of the UDS directly on the same underlying scale in terms of relative cutoffs and overall reliability and agreement.

4.1 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 synched to 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 multi-rater model employed here, by contrast, allows us to estimate the ten raters' score cutoff points (the model's γ parameters) along a uniform scale, making direct comparisons across scores possible.

[Figure 4 about here.]

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 post-war 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, 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 per cent highest posterior density interval. For example, the model indicates that there is a 95 per cent 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 level. These large error bars can be a result of rater inconsistency, a paucity of observations at a particular rating level, or both.

The estimation of these cutoffs allows us to answer a number of substantively interesting questions. The first thing to 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 the various judges are able to discriminate between cases effectively and exhibit sufficient reliability for levels on their scores to have meaning. Nonetheless, there are a number of notable discrimination problems. Take, for example, the meaning of the middle scores (from approximately -1 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 extremely close together, and in fact overlap considerably. There are simply too few cases 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 ordinal 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 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 stair-step fashion, exhibiting relatively uniform distances between cutoffs across the entire scale. On the other hand, Polity’s scores, even those at the ends of the scale, 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 & Ward 1997, Treier & Jackman 2007). In general, our component scales demonstrate reasonably consistent cutoff placement although virtually every rater exhibits some inconsistency at the high and low ends of the democracy scale. Cutoff consistency is important because researchers treat democracy scores as interval—or sometimes continuous—measures in their analyses and large nonlinearities in cutoff placement can potentially bias results.¹⁴

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 ordinal measures. A number of theories rely on the presence of democracy, not on level of democracy, to explain phenomena, requiring such an approach.¹⁵ A cutoff used by a variety of IR scholars is to code nations with democracy scores greater than 6 or 7 on the combined Polity scale as democracies and all other countries

¹⁴It is also important to note that the manner in which we discretize the continuous measures (See footnote 8) tends to obscure consistency issues in those scales. For example, although cutoffs based on quantiles of Arat’s measure are relatively even, the quantiles fall irregularly across Arat’s continuous ratings. Previous experiments fitting the model with evenly spaced cutoffs on the continuous scales provided evidence of minor consistency problems in Hadenius’ scale and substantial consistency issues with Arat’s scale. Vanhanen’s scale exhibited substantial inconsistency at high levels of democracy while Bollen’s cutoffs demonstrated consistent spacing.

¹⁵Findings of the democratic peace literature spring to mind.

as non-democracies (Ray 2000), while 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 in fact consistent with the PACL definition of democracy: the PACL cutoff lies somewhere between 5 and 6 on Polity’s scale.

4.2 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.¹⁶ The multi-rater 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 per cent credible intervals; high variance means more errors and less reliability and vice versa.

[Figure 5 about here.]

The overall picture from the reliability comparison is encouraging for those scholars who argue for democracy scale agnosticism, at least among the most popular measures (Adcock & Collier 2001). In the aggregate, the big three raters—Freedom House, PACL, and Polity—generally have moderate to high reliability when compared to the remaining measures. The error variances of Freedom House and Polity cannot even be confidently distinguished from one another. PACL, furthermore, is one of the most reliable raters in the UDS, making

¹⁶For example, Bollen (1980) places measure reliability at the center of his analysis

relatively few mistakes when categorizing countries and is statistically more reliable than either of the other two popular measures.¹⁷ 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 in by any of the three raters.

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.¹⁸ Furthermore, the most reliable raters in the sample are smaller, more focused, projects like Coppedge & Reinicke’s (1991) Polyarchy, Hadenius’ (1992) scale and especially Bowman, Lehoucq & Mahoney’s (2005) measure of democracy in Central America. 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.

¹⁷Note that PACL’s low error variance is not simply a function of its limited categorization scheme. A low error variance estimate does not merely indicate that PACL makes few misclassification errors but that all the raters generally consider the cases it does mis-classify to be close to PACL’s democracy-autocracy cutoff. Therefore, the low error variance estimate for PACL indicates that, while not technically based on a continuous conception of democracy, PACL considers the countries generally judged more democratic (authoritarian) by other raters less likely to be authoritarian (democratic) in a manner that decreases smoothly in a country’s distance from PACL’s cutoff (as defined in terms of the consensus measure). Thus, while PACL is explicitly not continuous in construction, it behaves in a manner consistent with a conceptualization of democracy as a continuous latent variable.

¹⁸In fact, there may be reason to think that Vanhanen is so unreliable because he is not measuring the same concept of democracy as the other measures do. His scale is an aggregation of voter turnout and the proportion of seats held by the largest party in the legislature that, while based on a standard minimalist definition of democracy, seems out of touch with the conceptualization of democracy captured by most raters. It is unclear however, whether Vanhanen’s measure is systematically biased or if his blunt instruments simply make his scores extremely unreliable, as we assume here.

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 & 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.

While 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 these 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.

4.3 Rater Differences

Throughout this paper, we have argued that, despite the high correlations between the ten 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 (Casper & Tufis 2002, Elkins 2000). 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.¹⁹ Based on data from our model, of the 3929 country-years when both Freedom House and Polity provide scores, 2380, or nearly 61 per cent, of the scores are statistically distinguishable from each other. Of the 4,698 country-years for which both Freedom House and PACL provide scores, 529, or about 11 per cent, of the scores are statistically distinguishable from each other. Finally, of the 6481 country-years when both PACL and Polity provide scores, 732, again about 11 per cent, of the scores are statistically distinguishable. Looking at the 3929 country-years when all three measures are present, only a measly 74, or less than 2 per cent, of the scores are different across all three measures. When looking at any two of these measures, however, as many as 50 per cent of the scores are discrepant.²⁰

It is important to put these findings in context; the majority of discrepancies, although

¹⁹We consider rater j 's perception of country-year i 's level of democracy statistically distinguishable from rater \hat{j} 's perception if either $\Pr(t_{ij} > t_{i\hat{j}}) > .95$ or $\Pr(t_{ij} < t_{i\hat{j}}) > .95$. Given $k = 1, \dots, K$ simulated draws from the joint posterior of the model parameters, we calculated $l_{ij\hat{j}} = \sum_k L(t_{ij}^{(k)}, t_{i\hat{j}}^{(k)})$ and $g_{ij\hat{j}} = \sum_k G(t_{ij}^{(k)}, t_{i\hat{j}}^{(k)})$ for every i , where L and G are indicator functions that return 1 if their first argument is less (greater) than their second argument and return 0 otherwise. The proportion of judge j 's ratings classified as distinguishable from judge \hat{j} 's equals

$$\frac{\sum_i G(l_{ij\hat{j}}, 0.95 \cdot n_{j\hat{j}}) + G(g_{ij\hat{j}}, 0.95 \cdot n_{j\hat{j}})}{n_{j\hat{j}}}$$

where $n_{j\hat{j}}$ is the number of country-years rated by both raters.

²⁰In 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.

statistically significant, are substantively small. Yet, there are cases where particular raters provide democracy estimates that are substantially at odds with the judgments of 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 especially unusual observation. Researchers intent on using an isolated democracy score in their research should use the UDS to identify such observations in their datasets 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 17 per cent of Freedom House's 4703 ratings have a 95 per cent chance of being strictly greater than or strictly less than the UDS composite score, while more than two per cent of PACL's 7457 scores meet this criterion,²¹ and over 15 per cent of Polity's 6577 scores are statistically distinguishable from the corresponding unified estimate.

5 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 & Jackman 2007). 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 error in empirical work on

²¹Again, these figures tend to grow with the number of rater cutoffs, all else equal.

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, it is possible to incorporate covariates in the multi-rater 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. Secondly, 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 multi-rater 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 & 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 & David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.
- Alvarez, Michael, José Antonio Cheibub, Fernando Limongi & Adam Pzeworski. 1996. "Classifying Political Regimes." *Studies in Comparative Political Development* 31(2):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(2):370–390.
- Bollen, Kenneth A. 2001. "Cross-National Indicators of Liberal Democracy, 1950-1990." 2nd ICPSR version. Chapel Hill, NC: University of North Carolina, 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2001. Retrieved from <http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/02532.xml>.
- Bollen, Kenneth A. & Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1):58–86.
- Bollen, Kenneth A. & Robert W. Jackman. 1989. "Democracy, Stability, and Dichotomies." *American Sociological Review* 54(4):612–621.
- Bowman, Kirk, Fabrice Lehoucq & James Mahoney. 2005. "Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America." *Comparative Political Studies* 38(8):939–970.
- Casper, Gretchen & Claudiu Tufis. 2002. "Correlation versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy using Highly Correlated Datasets." *Political Analysis* 11(2):1–11.
- Cheibub, Jose & Jennifer Gandhi. 2004. Classifying Political Regimes: A Sixfold Classification of Democracies and Dictatorships. In *Annual Meeting of the American Political Science Association*. Chicago, IL: .
- Collier, David & Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science* 2:537–565.

- Coppedge, Michael. 2002. "Democracy and Dimensions: Comments on Munck and Verkuilen." *Comparative Political Studies* 35(2):35–39.
- Coppedge, Michael & Wolfgang H. Reinicke. 1991. Measuring Polyarchy. In *On Measuring Democracy: Its Consequences and Concomitants*, ed. Alex Inkeles. New Brunswick, NJ: Transaction pp. 47–68.
- Dahl, Robert A. 1972. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.
- Elkins, Zachary. 2000. "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations." *American Journal of Political Science* 44(2):287–294.
- Freedom House. 2007. "Freedom in the World." <http://www.freedomhouse.org>.
- Gasiorowski, Mark J. 1996. "An Overview of the Political Regime Change Dataset." *Comparative Political Studies* 29(4):469–483.
- Gleditsch, Kristian S. & Michael D. Ward. 1997. "Double Take: A Reexamination of Democracy and Autocracy in Modern Polities." *The Journal of Conflict Resolution* 41(3):361–368.
- Hadenius, Axel. 1992. *Democracy and Development*. Cambridge: Cambridge University Press.
- Johnson, Valen E. 1996. "On Bayesian Analysis of Multirater Ordinal Data: An Application to Automated Essay Grading." *Journal of the American Statistical Association* 91(433):42–51.
- Johnson, Valen E. & James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer.
- Marshall, Monty G., Keith Jaggers & Ted Robert Gurr. 2006. "Polity IV: Political Regime Characteristics and Transitions, 1800-2004." <http://www.cidcm.umd.edu/polity/>.

- Munck, Gerardo L. & Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35(1):5–34.
- Przeworski, Adam, Michael Alvarez, José Antonio Cheibub & Fernando Limongi. 2000. *Democracy and Development: Political Regimes and Economic Well-being in the World, 1950–1990*. Cambridge: Cambridge University Press. Data available at <http://www.ssc.upenn.edu/~cheibub/data/Default.htm>.
- Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12(4):338–353.
- Ray, James Lee. 2000. Democracy: On the Level(s), Does Democracy Correlate with Peace? In *What Do We Know About War?*, ed. John Vasquez. Lanham: Rowman and Littlefield pp. 299–316.
- Treier, Shawn & Simon Jackman. 2007. "Democracy as a Latent Variable." <http://jackman.stanford.edu/papers/>. (Last accessed November, 2007).
- Vanhanen, Tatu. 2003. *Democratization: A Comparative Analysis of 170 Countries*. New York: Routledge.

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	60–155	1946–2000	-10–10	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	38–97	1946–1992	1–4	Competitiveness, Inclusiveness, and Political Liberties
Vanhanen	41–157	1946–2000	0.00–54.00	Competition and Participation

Table 2: Democracy Measure Correlation Matrix

	Arat	BLM	Bollen	F. House	Hadenius	PACL	Polity	Polyarchy	PRC	Vanhanen
Arat										
BLM	0.604 (175)									
Bollen	0.906 (480)	0.742 (25)								
F. House	0.857 (1511)	0.685 (145)	0.938 (152)							
Hadenius	0.649 (5)	0.649 (5)	0.948 (125)							
PACL	0.810 (3830)	0.692 (275)	0.816 (499)	0.828 (4698)	0.838 (128)					
Polity	0.853 (3569)	0.827 (263)	0.881 (465)	0.905 (3929)	0.913 (108)	0.856 (6481)				
Polyarchy	0.743 (10)	0.743 (10)	0.905 (345)	0.905 (345)	0.905 (108)	0.805 (350)	0.902 (286)			
PRC	0.688 (2482)	0.881 (235)	0.721 (325)	0.776 (1936)	0.843 (94)	0.753 (3662)	0.811 (3575)	0.759 (95)		
Vanhanen	0.771 (2161)	0.691 (202)	0.761 (304)	0.730 (3006)	0.689 (70)	0.682 (4661)	0.751 (4051)	0.685 (246)	0.603 (1675)	

Figure 1: Standardized Measures of Democracy over Time

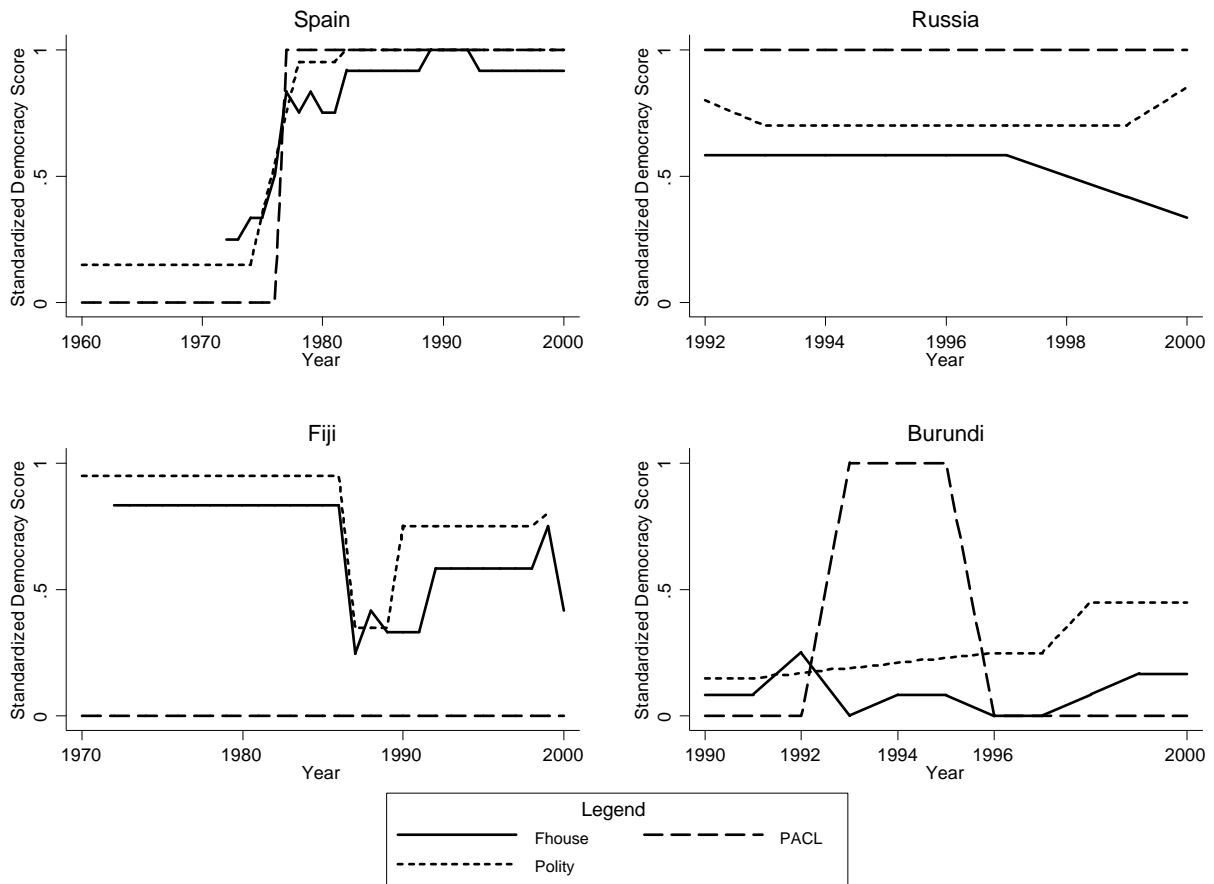


Figure 2: Unified Democracy Scores for 2000

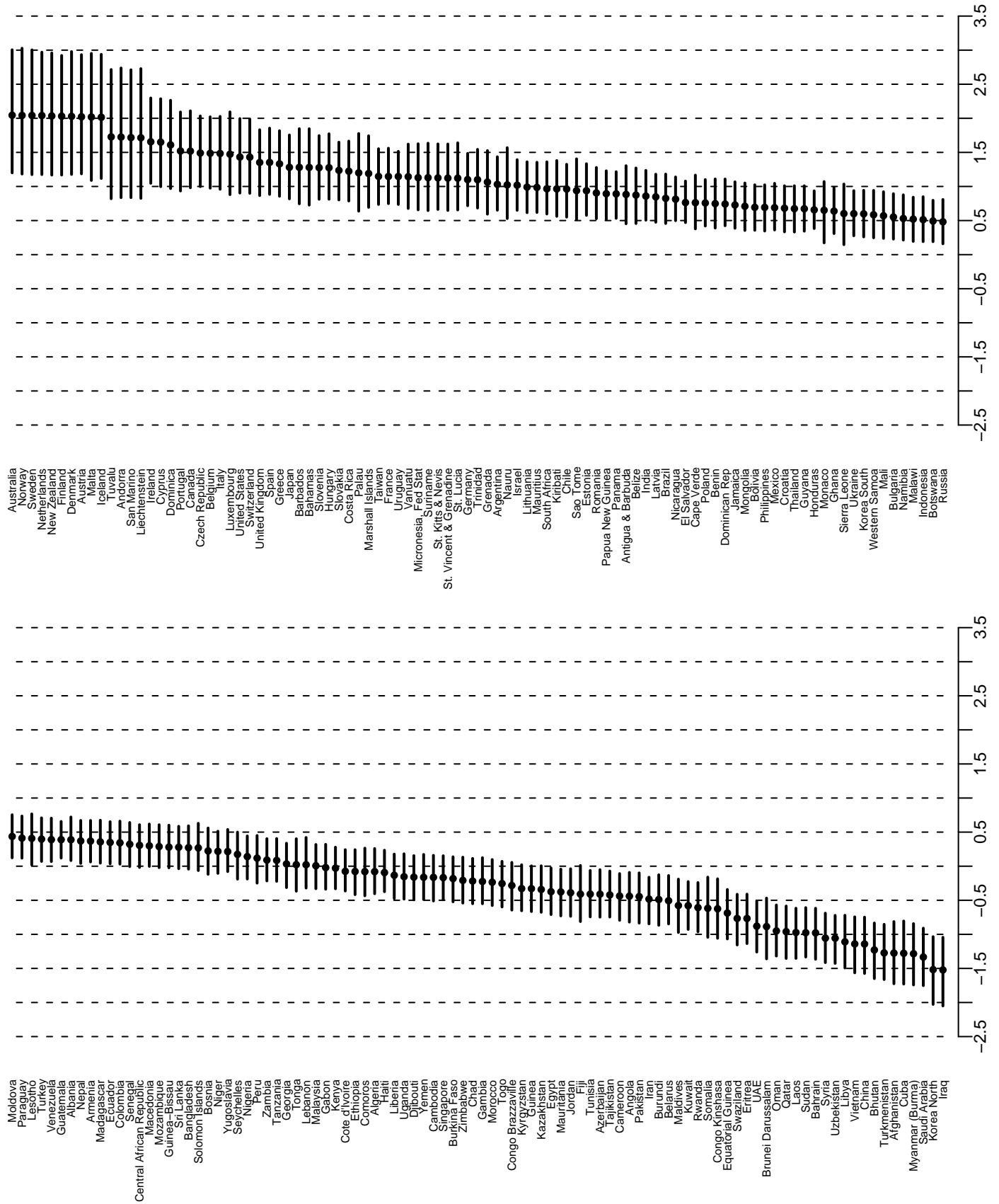


Figure 3: Observations and Mean UDS Standard Deviation by Year

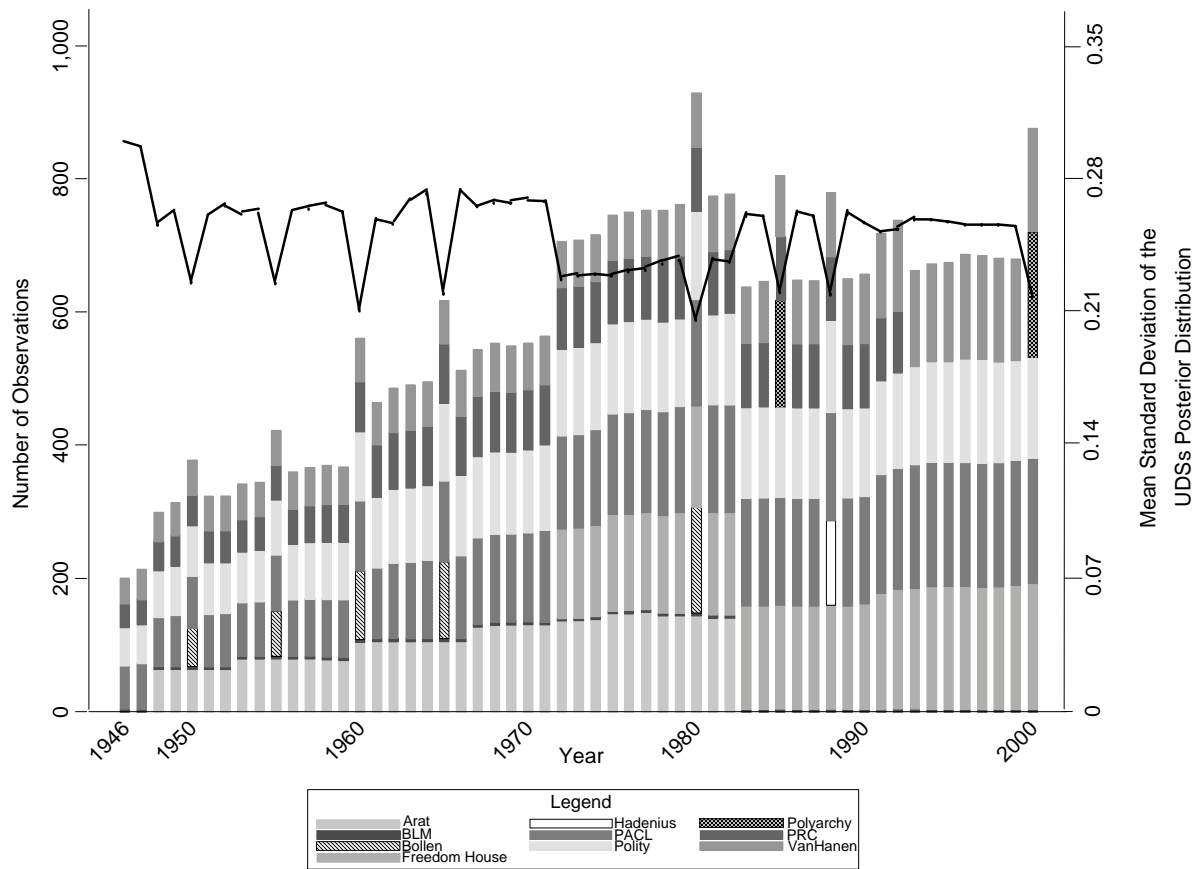


Figure 4: Democracy Measure Rating Cutoffs

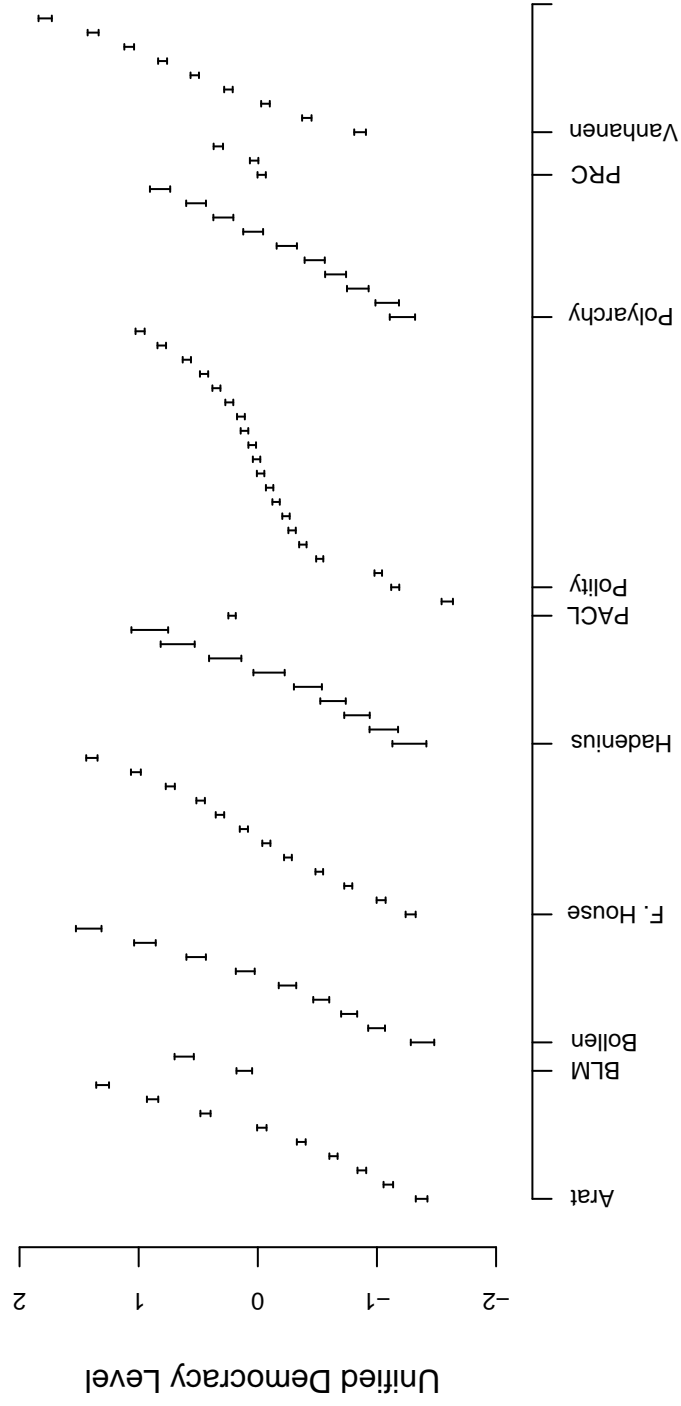


Figure 5: Democracy Measure Error Variance

