Democracy as a Latent Variable

Shawn Treier University of Minnesota
Simon Jackman Stanford University

We apply formal, statistical measurement models to the Polity indicators, used widely in studies of international relations to measure democracy. In so doing, we make explicit the hitherto implicit assumptions underlying scales built using the Polity indicators. Modeling democracy as a latent variable allows us to assess the "noise" (measurement error) in the resulting measure. We show that this measurement error is considerable and has substantive consequences when using a measure of democracy as an independent variable in cross-national statistical analyses. Our analysis suggests that skepticism as to the precision of the Polity democracy scale is well founded and that many researchers have been overly sanguine about the properties of the Polity democracy scale in applied statistical work.

Social and political theories often refer to constructs that cannot be observed directly. Examples include the ideological dispositions of survey respondents (e.g., Erikson 1990), legislators (Clinton, Jackman, and Rivers 2004), judges (Martin and Quinn 2002), or political parties (Huber and Inglehart 1995); the quantitative, verbal, and analytic abilities of applicants to graduate school (e.g., the GREs); locations in an abstract, latent space used to represent relations in a social network (Hoff, Raftery, and Handcock 2002); levels of support for political candidates over the course of an election campaign (e.g., Green, Gerber, and De Boef 1999). In each instance, the available data are manifestations of the latent quantity (*indicators*) and the inferential problem can be stated as follows: given observable data y, what should we believe about latent quantities x?

A prominent example—and the subject of this article—is constructing country-level measures of democracy. Measures of democracy are used extensively in empirical work on the "democratic peace" and economic development. As we outline below, even a casual survey of this literature reveals an uneasiness with extant measures of democracy. Various indicators of democracy are combined in seemingly arbitrary ways, without any formal or explicit justification of the procedure used to map from indicators to the derived measure. We offer a number of improvements on this procedure. We use a measurement model to derive a rule for combining the information in indicators of democracy so as to produce a score for each country-year observation. We contrast the measures of democracy obtained from this model-based approach with extant measures. We quantify the uncertainty in our measures of democracy, which is considerable, arising from the facts that (a) we have relatively few indicators of democracy available for analysis; and (b) we have no strong prior theoretical reasons to believe any one indicator is a better (or worse) indicator than any other indicator. We then demonstrate that this measurement uncertainty can be consequential, making it difficult to draw reliable inferences about the impact of democracy on outcomes of substantive interest such as interstate conflict. We close with an assessment of the circumstances in which this measurement uncertainty is likely to be consequential and when it might be ignored.

Measuring Democracy

A tremendous amount of time and effort has been (and continues to be) devoted to measuring democracy, or

Shawn Treier is assistant professor of political science, University of Minnesota, 1414 Social Sciences Building, 267 19th Avenue South, Minneapolis, MN 55455 (satreier@umn.edu). Simon Jackman is professor of political science, Stanford University, 616 Serra St., Encina Hall West, Room 100, Stanford, CA 94305-6044 (jackman@stanford.edu).

Earlier versions of this work were presented at the 2003 annual meeting of the Midwestern Political Science Association, 2003 annual meeting of the Society for Political Methodology, and the 2003 annual meeting of the American Political Science Association. We thank Larry Bartels, Neal Beck, Jon Bendor, Bruce Bueno De Mesquita, Alberto Diaz, Jim Fearon, William Jacoby, Steve Krasner, David Laitin, Jeffrey Lewis, Nikolay Marinov, Andrew Martin, Doug Rivers, and Mike Tomz for useful comments and references. Errors and omissions remain our own responsibility.

American Journal of Political Science, Vol. 52, No. 1, January 2008, Pp. 201–217

 ISSN 0092-5853

more specifically, assigning annual scores to countries on specific indicators of democracy. But what, exactly, do these scores tap? What is the nature of the underlying latent construct, democracy? Investigations of these foundational questions appear to be rare. As Munck and Verkuilen lament, "... with a few notable exceptions, quantitative researchers have paid sparse attention to the quality of the data on democracy that they analyze ... To a large extent, problems of causal inference have overshadowed the equally important problems of conceptualization and measurement" (2002, 5–6). There are some important exceptions to this general observation. For instance, Bollen (1993) demonstrates validity problems with several additive indices of democracy caused by rater biases in the original assignment of values to the indicators. Gleditsch and Ward (1997) extensively describe the coarseness of the Polity index, casting doubt on the level of measurement of the scale. And Coppedge and Reinicke (1991) take greater care than most researchers in evaluating the assumptions of the additive index, constructing a Guttman scale from a set of indicators of polyarchy.

Munck and Verkuilen (2002) detail many of the deficiencies of current methods. We concur with their assessment and their conclusions: i.e., a good measure of democracy should identify the appropriate attributes that constitute democracy, each represented by multiple observed indicators; have a well-conceived view of the appropriate level of measurement for the indicators and the resulting scale; and should properly aggregate the indicators into a scale without loss of information. Most applications are deficient on at least one of these counts. For instance, in selecting the number of indicators, researchers tend toward "minimalist" definitions using only a few variables, which are often insufficient to separate out different gradations of democracy (Munck and Verkuilen 2002, 10–12). Some researchers even operationalize democracy with a single indicator, seeing approaches based on multiple indicators as unnecessarily complicated (e.g., Gasiorowski 1996). However, the hope that a solitary indicator circumvents these measurement issues is illusory; indeed, most scholars agree that democracy is multifaceted, and hence not well characterized by a single indicator.

Generating Scores from Indicators: The Problem of Aggregation

Among scholars who operationalize democracy via multiple indicators, there is no agreement regarding how one should *aggregate* the information in the indicators, a data reduction task whereby we assign a score to each country-year observation, given the scores on the indicator variables for that country-year. The democracy scores provided by the well-known Polity data set (Marshall and Jaggers 2002b) are combinations of indicators with an a priori specified weighting scheme. The apparent arbitrariness of the weighting/scoring scheme aside, the particular scoring rule used in Polity appears to discard much of the variation in the indicators: e.g., as Gleditsch and Ward (1997) observe, many different coding patterns across the Polity indicators are assigned the same Polity score, generating "lumpiness" in the distribution of Polity scores. Many researchers, apparently concerned by this feature of the Polity scores, reduce the Polity scores into three classifications: autocracy, anoncracy, and democracy (variations on the tripartite classification of Marshall and Jaggers 2002b, 32–33). This variation in extant measurement procedures suggests that there seems to be no settled method for aggregating indicators of democracy, or for evaluating justifications of these rules. Even some careful, rigorous investigations ignore the issue of aggregation/scoring. For instance, in a study of rater bias in coding indicators, Bollen (1993) takes the resulting additive indices as given and does not examine the question of how the indicators are aggregated to form democracy scales. Gleditsch and Ward (1997) describe many problems with Polity scores that are symptomatic of the arbitrary aggregation scheme, but do not investigate how one could improve the resulting scale. Only Coppedge and Reinicke (1991) deal directly with the aggregation issue, but their study is limited by the restrictive and deterministic assumptions of Guttman scaling.[1]

The Consequences of Measurement Error Are Seldom Acknowledged

A final deficiency is that scholars who either create or use measures of democracy seldom confront the issue of measurement error. That is, quite aside from the arbitrariness of an ad hoc aggregation rule, almost all simple aggregations presume a completely deterministic and perfect measurement process, ignoring that each of these indicators is an imperfect representation of democracy. As Bollen observes, "it is worthwhile to point out that in the typical multiple regression model, researchers assume that their democracy measures contain no random or systematic measurement error" (1993, 1218). Since most of the indicators Bollen considers are measured with error, any composite index must also be measured with error. Thus whenever democracy appears as an explanatory

[1]The limited applicability of Guttman scaling is apparent from the 20% of observations for which the response pattern is inconsistent with the scale (Munck and Verkuilen 2002, 23).

variable in empirical work, there is an (almost always ignored) errors-in-variables problem, potentially invalidating the substantive conclusions of these studies. The consequences of measurement error in regression models are well known, yet worth briefly repeating.[2] There are no adverse effects when the dependent variable is measured with error (the additional error is subsumed in the regression error), yet quite consequential effects when one of the independent variables is measured imperfectly. Estimated slopes are biased and inconsistent. With only one poorly measured variable (e.g., the democracy index), the coefficient on that variable is attenuated, while the others are biased and inconsistent in unknown direction and magnitude.[3]

The approach we present below explicitly confronts the fact that like any latent variable, democracy is measured with error. We show how a recent study of civil wars warrants reassessment in light of the measurement error inherent in democracy. Our contribution is to show that there are principled, statistical methods for using indicators of democracy to arrive at measures of regimes. To recapitulate, our position is that democracy is a latent variable, and cannot be measured directly, but that indicators of democracy (of varying degrees of fidelity) are available. Thus, measuring democracy is an *inferential* problem. Specifically, we address two questions: (1) how to best aggregate the information in the indicators, and (2) how to ensure that whatever uncertainty exists in the resulting measure or classification of democracy propagates into subsequent statistical uses of the measure. In particular, we show how to guard against a false sense of security when using measures of democracy as an independent variable in a regression analysis. Inferences about the effects of democracy on some outcome of interest should reflect the fact that democracy is a latent variable, measured via a limited number of imperfect indicators.

The Polity Data

Many different collections of indicators of democracy have been employed at one time or another in studies of international relations and comparative politics (see the enumeration in Munck and Verkuilen 2002). We base our empirical analysis on the extensively used measures from the Polity Project (Marshall and Jaggers 2002b). Polity IV covers the period of 1800–2000 for some 184 countries, for a total of 13,941 country-years;[4] more important, all of the indicators used to construct the aggregate measure are accessible and well documented, unlike some alternative measures. The summary measure used widely in empirical applications is a country-year's "Polity score," ranging from −10 to 10, created from five expert-coded categorical indicators: (1) Competitiveness of Executive Recruitment (*Xrcomp*), (2) Openness of Executive Recruitment (*Xropen*), (3) Executive Constraints/Decision Rules (*Xconst*), (4) Regulation of Participation (*Parreg*), (5) Competitiveness of Participation (*Parcomp*). Table 1, adapted from Marshall et al. (2002), illustrates the contribution of each value of the indicators to the Polity score.

A sixth variable, Regulation of Executive Recruitment (*Xrreg*), is not used directly in the calculation of the Polity score, but affects the coding rules for the other indicators. The six indicators sort into three categories: executive recruitment (XR), executive constraint (XCONST), and political participation (PAR), which define the alternative "concept variables" (*Exrec*, *Exconst*, and *Polcomp*). Exrec is constructed from Xrreg, Xrcomp, and Xropen; Polcomp is defined by Parreg and Polcomp; while Exconst is identical to Xconst. Additional information on the dataset is available in Marshall and Jaggers (2002a), but two aspects of the data are worth mentioning here. First, all of the indicators are ordinal except Xropen. Xropen = 4 ("Open") has two different Polity contributions, depending on the value of Xrcomp. Despite being the highest category, "Open" polities are not necessarily electoral democracies, but include polities that chose chief executives by elite designation; thus, the original coding is nominal. In our discussion, the category "Open" has been split into two categories, "Open, Election" (4^{E}) and "Open, No Election" (4^{NE}). Second, using either the six components or the three concept variables returns the same Polity score; there is no loss of information in moving from six indicators to the three concept variables.[5]

[2] For a summary of basic results, see any standard econometrics text (e.g., Greene 2003, 83–86). Book-length treatments are provided by Fuller (1987) and Wansbeek and Meijer (2000).

[3] The invalid aggregation of multiple indicators further complicates the likely consequences. Standard results assume that the imperfect measure X is a valid estimate of the true measure ξ, i.e., $E(X) = \xi$. If these measures are improperly aggregrated, the measure may contain substantial systematic bias ($E(X) \neq \xi$). Even in the simplest case (linear regression with one regressor), the direction of the bias is unknown; the estimates may be attenuated or *larger* than the true effect (Carroll, Ruppert, and Stefanski 1995).

[4] All observations with missing indicators (coded −66, −77, or −88) are excluded from the analysis.

[5] There are a few exceptions. Some infrequent patterns of (Xrreg, Xropen, Xrcomp) were excluded from the definition of Exrec; similarly, infrequent patterns of (Parreg, Parcomp) were excluded from the definition of Polcomp. Country-years with these patterns, despite having fully observed subindices and corresponding Polity scores, were not assigned a value for Exrec or Polcomp. To correct these apparent omissions, we redefined Exrec and Polcomp, assigning these handful of cases to separate categories in the concept

Table 1 Description of Polity Coding Rules

Indicators	Values	Democracy	Autocracy	Polity	Implied Order
PARCOMP:					
Competitive	5	3	0	3	6
Transitional	4	2	0	2	5
Factional	3	1	0	1	4
Restricted	2	0	1	−1	2
Suppressed	1	0	2	−2	1
Not applicable	0	0	0	0	3
PARREG:					
Regulated	5	0	0	0	3
Multiple Identity	2	0	0	0	3
Sectarian	3	0	1	−1	2
Restricted	4	0	2	−2	1
Unregulated	1	0	0	0	3
XRCOMP:					
Election	3	2	0	2	4
Transitional	2	1	0	1	3
Selection	1	0	2	−2	1
Unregulated	0	0	0	0	2
XROPEN:					
Open ("Election")	4	1	0	1	6
Dual: Hereditary and Election	3	1	0	1	5
Dual: Hereditary and Designation	2	0	1	−1	2
Closed	1	0	1	−1	1
Unregulated	0	0	0	0	4
Open ("No Elections")	4	0	0	0	3
XCONST:					
Parity or Subordination	7	4	0	4	7
Intermediate 1	6	3	0	3	6
Substantial	5	2	0	2	5
Intermediate 2	4	1	0	1	4
Slight Moderation	3	0	1	−1	3
Intermediate 3	2	0	2	−2	2
Unlimited Power	1	0	3	−3	1

Note: Adapted from Table 1 in Marshall et al. (2002).

An Ordinal Item-Response Model for the Polity Indicators

The aggregation or scoring rule for the Polity index is extremely simple; usually a one category increase on any one of the ordinal indicators generates a unit increase in the Polity score. But is this the most appropriate aggregation rule for these indicators? Can the Polity indicators be treated as interval measures? Should moving from 1 to 2 on indicator *j* have the same contribution to the resulting measure of democracy as, say, moving from 3 to 4 on the same indicator? Moreover, do all indicators tap the latent construct (democracy) equally well? That is, should a move from 1 to 2 on indicator *j* have the same impact as an increase from 1 to 2 on indicator *k*? In short, to what extent is the aggregation rule employed by Polity supported by the data?

variable, and established the ordering of the new categories based on information from the components and the contribution to the Polity score. Consequently, in this analysis, Exrec and Polcomp have 11 and 12 categories (eight and 10 in the original data).

We address these issues with the following statistical model. We treat democracy as a latent, continuous variable. The ordinal Polity IV indicators for each country-year are modeled as functions of the unobserved level of democracy, via the following ordinal item-response model. Let $i = 1, \ldots, n$ index country-years and $j = 1, \ldots, m$ index the Polity indicators. Let $k = 1, \ldots, K_j$ index the (ordered) response categories for item j. Then our model is

$$\begin{aligned}
\Pr(y_{ij} = 1) &= F(\tau_{j1} - \mu_{ij}) \\
&\vdots \quad \vdots \\
\Pr(y_{ij} = k) &= F(\tau_{jk} - \mu_{ij}) - F(\tau_{j,k-1} - \mu_{ij}) \\
&\vdots \quad \vdots \\
\Pr(y_{ij} = K_j) &= 1 - F(\tau_{j,K_j-1} - \mu_{ij})
\end{aligned}$$

where $\mu_{ij} = x_i\beta_j$, x_i is the latent level of democracy in country-year i, y_{ij} is the i-th country-year's score on indicator j, and $F(\cdot)$ is a function mapping from the real line to the unit probability interval, defined here as the logistic CDF $F(z) = 1/(1 + \exp(-z))$. The slope parameter β_j is the *item discrimination parameter*, tapping the extent to which variation in the scores on the latent concepts generates different response probabilities. These parameters are referred to as item discrimination parameters because if item j does not help us distinguish among countries with different levels of democracy (x_i), then β_j will be indistinguishable from zero. $\boldsymbol{\tau}_j$ is a vector of unobserved thresholds for item j, of length $K_j - 1$, that follow an ordering constraint implied by the ordering of the responses, i.e., $\tau_{ja} < \tau_{jb}, \forall a < b, \forall j$.

For the uninitiated, it may help to note that item-response models are analogous to factor analysis models, with item-discrimination parameters analogous to factor loadings; the similarities between the two models are elaborated in Takane and de Leeuw (1987) and Reckase (1997). That said, there are some important differences between factor analysis (as conventionally implemented) and our fully Bayesian, item-response approach. In a Bayesian analysis, the goal is to characterize the joint posterior density of all parameters in the analysis. This means that the latent variables x are estimable and subject to inference just like any other parameters in the model. Thus, the latent variables have a different status in an item-response model than in conventional factor analysis. The typical implementation of factor analysis is as a model for the covariance matrix of the indicators (and not for the indicators per se), without the identifying restrictions necessary to uniquely recover factor scores, and hence the multiple proposals for obtaining factor scores conditional on estimates of a factor structure (e.g., Mardia, Kent, and Bibby 1979, sec. 9.7). Contrast the item-response approach in which the observed indicators—the "response" part of "item-response"—are modeled directly as functions of the latent variables. Incorporating the latent variables as parameters to be estimated comes at some cost: the number of parameters to be estimated is now potentially massive (i.e., one latent trait per country-year), but with the desktop computing resources now available to social scientists, estimating a fully Bayesian ordinal IRT (item-response theory) model for the Polity data poses no great challenge.

Key Assumption: Local Independence

An important assumption underlying both IRT models and factor analytic models is *local independence*, the property that the indicators y_{ij} are conditionally independent given the latent variable x_i: i.e.,

$$\begin{aligned}
&\Pr(y_{i1}, y_{i2}, \ldots, y_{im} \mid x_i) \\
&\quad = \Pr(y_{i1} \mid x_i)\Pr(y_{i2} \mid x_i) \cdots \Pr(y_{im} \mid x_i).
\end{aligned}$$

If local independence holds, the correlation between any two observed variables is due solely to the relationship of each variable to the unobserved latent factor; conditioning on that factor, the two variables are independent. There are numerous ways local independence might be violated. In particular, a violation of local independence occurs in educational testing when knowledge of the answer on previous questions is necessary for a correct answer on the current question (e.g., the first question on a statistics exam requires the calculation of the mean, the second question requires the calculation of the variance).

The Polity indicators are patently not locally independent. This is clear from the sparseness of the cross-tabulations of the Polity indicators in Table 2. Certainly, we do not expect patterns of complete independence in the table; Table 2 does not condition on the unobserved level of democracy, and since both indicators are presumably related to democracy, we should observe a relationship. Nevertheless, of the 24 possible combinations of values between Xrcomp and Xropen, only eight are observed.[6] Not observing all possible combinations is hardly unusual and of itself does not constitute a violation of local independence. But in this case, the proliferation and pattern of empty cells in Table 2 clearly identifies a pattern of dependence in the coding of the two variables. Most egregiously, if Xrcomp is 0, then Xropen is *always* 0. Irrespective of the underlying level of democracy, if we observe Xrcomp = 0,

[6] Gleditsch and Ward (1997) also provide evidence of the extreme overlap between the Polity indicators.

TABLE 2 Local Dependence in XR and PAR Polity Indicators

	Xropen					
Xrcomp	**1**	**2**	**4^{NE}**	**0**	**3**	**4^{E}**
1	2487	1175	3782	0	25	0
0	0	0	0	1600	0	0
2	0	0	0	0	130	667
3	0	0	0	0	0	4075
	Parcomp					
Parreg	**1**	**2**	**0**	**3**	**4**	**5**
4	3878	1811	0	0	0	0
3	0	299	0	3509	76	0
1	0	0	487	0	10	0
2	0	0	96	740	583	0
5	0	0	0	0	116	2336

Notes: Rows and columns have been reordered so as to highlight the dependencies between the two pairs of indicators.

then logically $\Pr(\text{Xropen} = 0) = 1$. The extreme sparseness in the first row and first column suggests a deterministic relationship rather than an unlucky random draw of cases. This dependency holds for every value of both variables of Xropen and Xrcomp, resulting in a nearly diagonal distribution of the cases through Table 2. A similar pattern emerges for Parreg and Parcomp, displayed in the lower half of Table 2.

A solution to the local independence problem is to combine the information from the six indicators, some of which are conditionally dependent, creating a smaller number of locally independent indicators. For Polity, the concept variables Exrec, Polcomp, and Exconst are three ordered indicators with 11, 12, and seven categories, which preserve the ordinal information in the six Polity indicators, yet can be considered locally independent. The resulting logistic IRT model is easily estimated via the simulation methods (Markov Chain Monte Carlo methods) described in the appendix.

Evaluating a Measure of Democracy

Table 3 presents the estimated discrimination parameters and thresholds for each item. All of the items discriminate well with respect to the latent trait, with Exconst (Executive Constraints) providing the highest discrimination, and Exrec (Executive Recruitment) with the smallest discrimination parameter. None of the indicators are unrelated to the latent trait, but there is variation in item discrimination: some of the Polity indicators tap the latent trait better than others, and any scale measure based on the Polity indicators ought to reflect this (as our measure does).

In addition to assuming equal importance of the three indicators, the Polity calculation imposes restrictive assumptions on the way movement within any given indicator contributed to the final score. In additive, linear scales (such as Polity), the ordered indicators are treated implicitly as interval measures, with the level of the underlying construct increasing linearly for every advancement to the next highest category, on any given indicator. This constitutes an extremely strong assumption, and one that is likely false, given the pattern of threshold estimates in Table 3. For instance, for the Polcomp indicator, the largest distance between thresholds occurs between categories 5 and 6, but these two categories have exactly the same Polity contribution. There are a few exceptions; for instance, our estimates suggest that collapsing 4 and 5, and 7 and 8 on Polcomp and 7 and 8 on Exrec is reasonable. Nevertheless, it is generally the case that the pattern of threshold estimates we obtain does not conform with the a priori specification of the Polity calculation.

In Figure 1 we compare (1) the posterior means of the latent traits from our ordinal IRT model; (2) factor scores from classical factor analysis, using "regression" scoring (ignoring the ordinal nature of the indicators); and (3) the Polity IV scores themselves. For clarity and simplicity, we restrict the comparison to the year 2000. Figure 1 shows the three pairwise scatterplots among the three candidate measures in a matrix of scatterplots; above the diagonal are the Pearson correlations among the three estimates. These correlations are all very large, and we might conclude that these measures of democracy are interchangeable. However, closer inspection reveals that at any given level of Polity, there is considerable variation in the range of corresponding latent traits found by the other two methods (our ordinal IRT model and classical factor analysis), or vice versa.

In particular, the S-shaped pattern in the mapping between Polity and our IRT estimates (bottom left panel of Figure 1) reflects the artificial "top-coding" in Polity: a score of 10 on Polity arises via the "maximum" response profile (11,12,7). This corresponds to an extremely high level on the latent scale underlying our IRT model, and the cluster of cases with this set of responses looks quite distinct from the rest of the data. Likewise at the bottom end of the Polity scale, there is considerable divergence with our estimates, due to the different weights our ordinal IRT model assigns to different indicators.

TABLE 3 Discrimination Parameters and Thresholds

Indicators	Discrimination Parameter		Thresholds	
Executive Recruitment (EXREC)	2.50 [2.42, 2.59]	τ_{11}	−2.66	[−2.73, −2.58]
		τ_{12}	−1.94	[−2.01, −1.88]
		τ_{13}	−1.93	[−2.00, −1.87]
		τ_{14}	−0.14	[−0.18, −0.09]
		τ_{15}	−0.11	[−0.16, −0.07]
		τ_{16}	0.67	[0.62, 0.72]
		τ_{17}	0.88	[0.83, 0.94]
		τ_{18}	0.99	[0.94, 1.05]
		τ_{19}	1.44	[1.38, 1.50]
		$\tau_{1,10}$	1.53	[1.47, 1.59]
Executive Constraints (EXCONST)	3.62 [3.47, 3.78]	τ_{21}	−1.96	[−2.04, −1.86]
		τ_{22}	−1.48	[−1.56, −1.40]
		τ_{23}	1.05	[1.00, 1.13]
		τ_{24}	1.28	[1.22, 1.36]
		τ_{25}	2.10	[2.02, 2.20]
		τ_{26}	2.44	[2.33, 2.54]
Political Competition (POLCOMP)	3.02 [2.92, 3.13]	τ_{31}	−2.02	[−2.10, −1.96]
		τ_{32}	−0.96	[−1.01, −0.90]
		τ_{33}	−0.80	[−0.86, −0.75]
		τ_{34}	−0.55	[−0.60, −0.50]
		τ_{35}	−0.50	[−0.55, −0.45]
		τ_{36}	1.85	[1.78, 1.92]
		τ_{37}	2.59	[2.52, 2.68]
		τ_{38}	2.67	[2.59, 2.77]
		τ_{39}	2.68	[2.60, 2.78]
		$\tau_{3,10}$	3.33	[3.22, 3.44]
		$\tau_{3,11}$	3.47	[3.36, 3.58]

Note: Posterior Means, with 95% Highest Posterior Density Intervals in brackets.

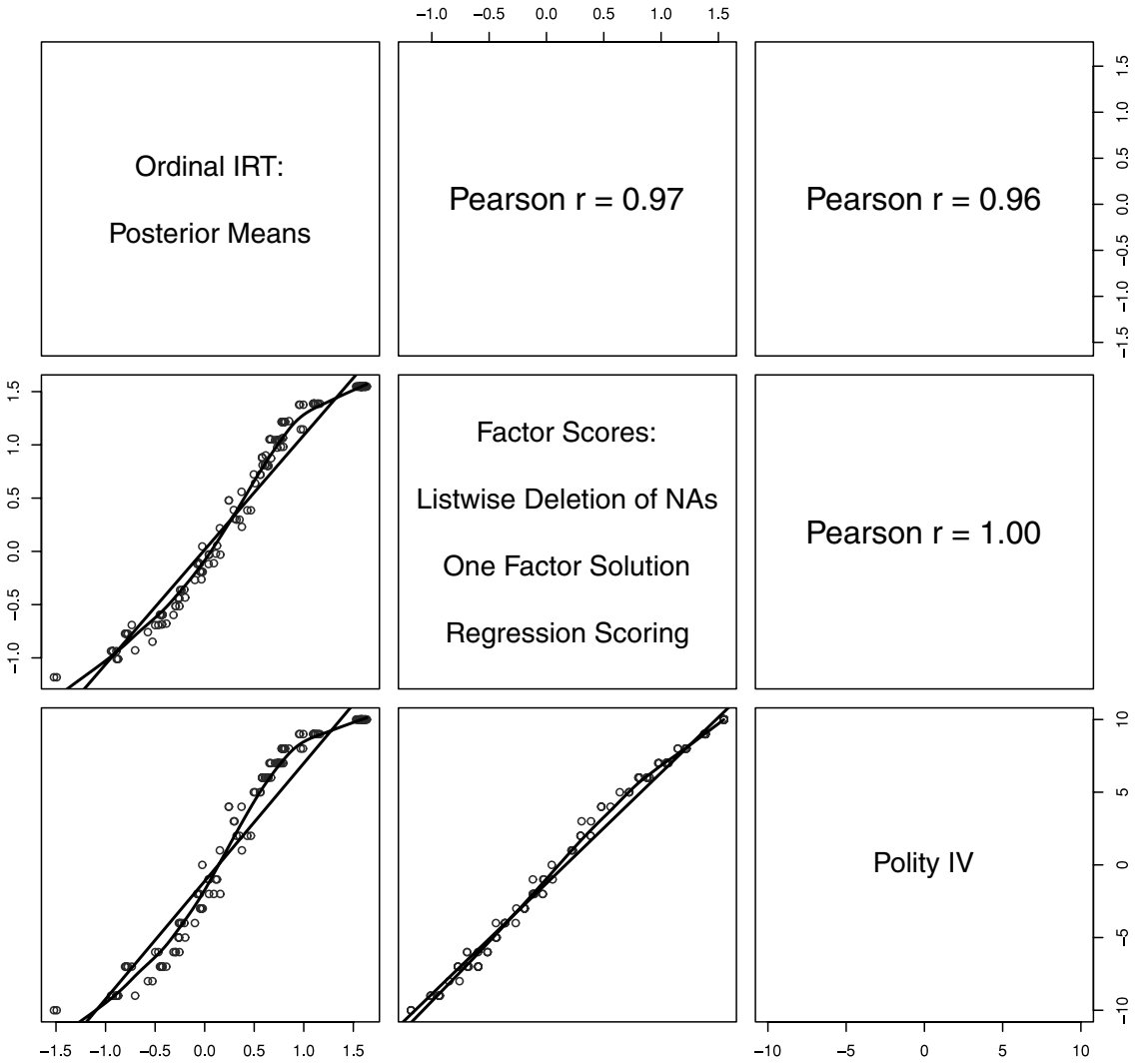
In Table 4 we closely inspect the dispersion of Polity scores within each decile of our estimated democracy scores (again, these are the posterior means of x_i in the ordinal IRT model). In just two deciles (the very top and the second to bottom) is the dispersion of Polity scores reasonably small. Elsewhere we find a wide range of Polity scores at any given level of the latent trait recovered by our model. So, although the correlation between our estimates and Polity is high, there is actually a surprising amount of divergence between the two approaches. For instance, a country-year that we would find, say, to lie in the 60–70% range on our democracy scale could have a Polity score between −2 and 6, a range that covers 40% of the 21-point Polity scale. Thus, if one were to treat our scores as "true scores," then the Polity scores look somewhat unreliable.

Assessing Measurement Error in the Latent Trait

Of course, a key feature of our approach is that we do not have to treat our estimates of latent democracy as "true scores": in our fully Bayesian analysis, we recover not just point estimates of latent democracy (means of the marginal posterior densities of latent levels of democracy, x_i) but also confidence intervals (quantiles of the marginal posterior densities). This makes it easy to compute and assess the measurement error in each country's latent level of democracy. Although we compute estimates covering the entire data period, for clarity and simplicity we concentrate primarily on the estimates for 2000 only.

Figure 2 displays the estimates for all 153 countries which received Polity IV codings in 2000. Unlike the

FIGURE 1 Comparison of Ordinal IRT Posterior Means, Factor Scores, and Polity, 2000



Note: Line for OLS local linear regression fits (span = 1/2, tri-cube kernel) are superimposed.

Polity scores, we are able to provide measures of uncertainty for each estimated latent score. The estimated scale ranges from Autocracy to Democracy from left to right. We summarize the marginal posterior density of each country-year's x_i with a point (the posterior mean) and a line covering a 95% highest posterior density (HPD) region.[7]

The striking feature of Figure 2 is that the measurement error increases in the extremes of the latent trait distribution. Countries that are estimated to have either extremely high or extremely low levels of democracy also have substantially larger levels of measurement error. This is actually a familiar result in IRT modeling. A country receiving an extremely high set of scores on the observed indicators is like the student in our classes who correctly answers all the questions on a test: we know that the student is at the top of the class, but until we see the student start to get items wrong, we cannot put an upper bound

[7]The following definition of an HPD is standard in the statistical literature (e.g., Bernardo and Smith 1994, 260). For a random variable $\theta \in \Omega$, a region $C \subseteq \Omega$ is a $100(1-\alpha)\%$ highest probability density region for θ if (1) $P(\theta \in C) = 1 - \alpha$; (2) $P(\theta_1) \geq P(\theta_2), \forall\, \theta_1 \in C, \theta_2 \notin C$. A $100\,(1-\alpha)\%$ HPD region for a random variable with a symmetric, unimodal density is obviously unique and symmetric around the mode of the density. In fact, if θ has a univariate normal density, an HPD is the same as a confidence interval around the mean.

TABLE 4 Comparison of Latent Trait Posterior Means and Polity Scores

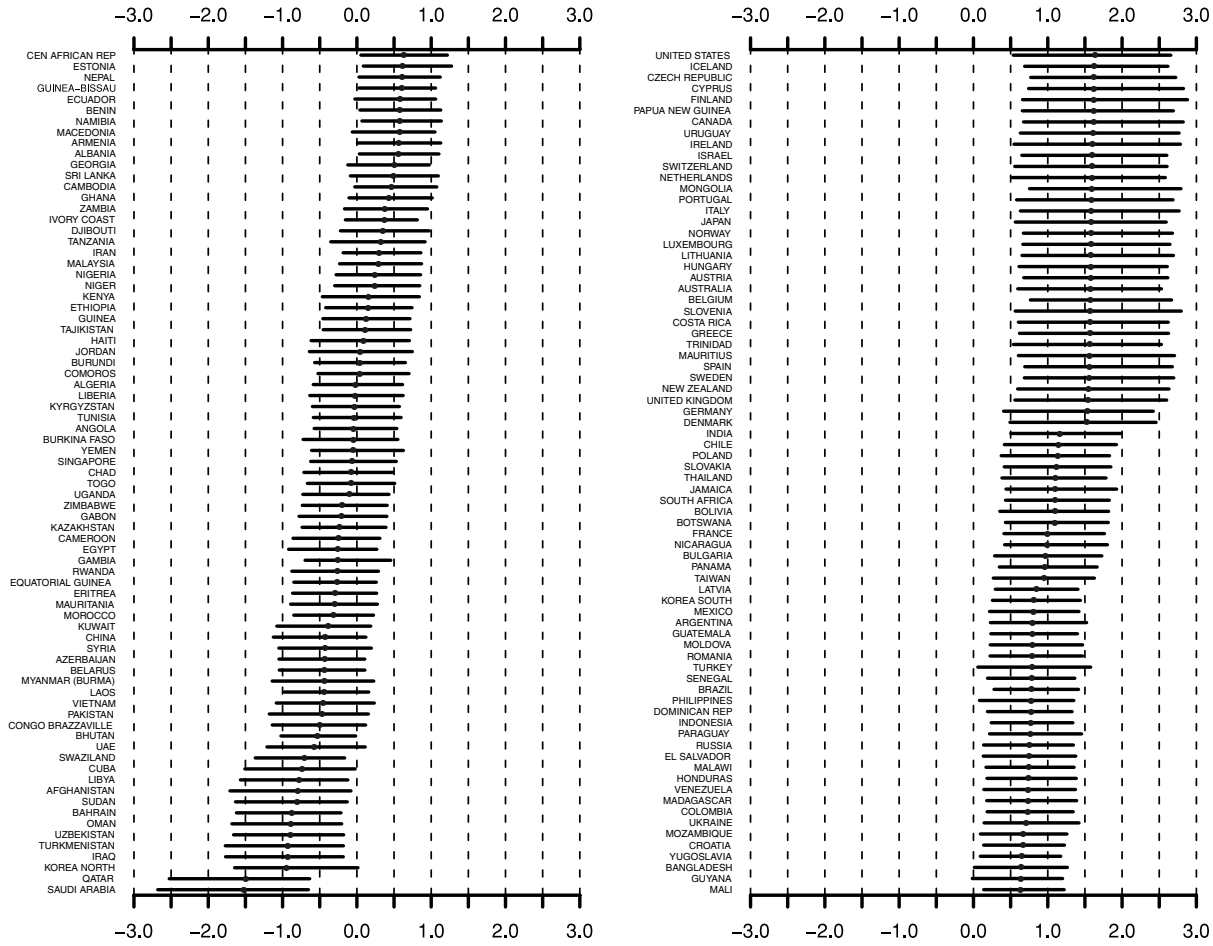
Ordinal IRT Model	Polity IV			
Latent Trait (Decile)	Median	2.5%	97.5%	Range
minimum–10%	−10	−10	−9	1
10%–20%	−9	−9	−6	3
20%–30%	−7	−9	−6	4
30%–40%	−7	−7	−5	5
40%–50%	−5	−7	−3	7
50%–60%	−3	−4	0	5
60%–70%	1	−1	5	8
70%–80%	6	3	8	6
80%–90%	9	8	10	6
90%–maximum	10	10	10	0

on our estimate of the student's ability. Countries assigned the maximum/minimum scores on the Polity indicators are like these students; we know that these countries are the most/least democratic in our data, but we do not get a precise estimate of the level of democracy in these countries.

Distinguishing Levels of Democracy

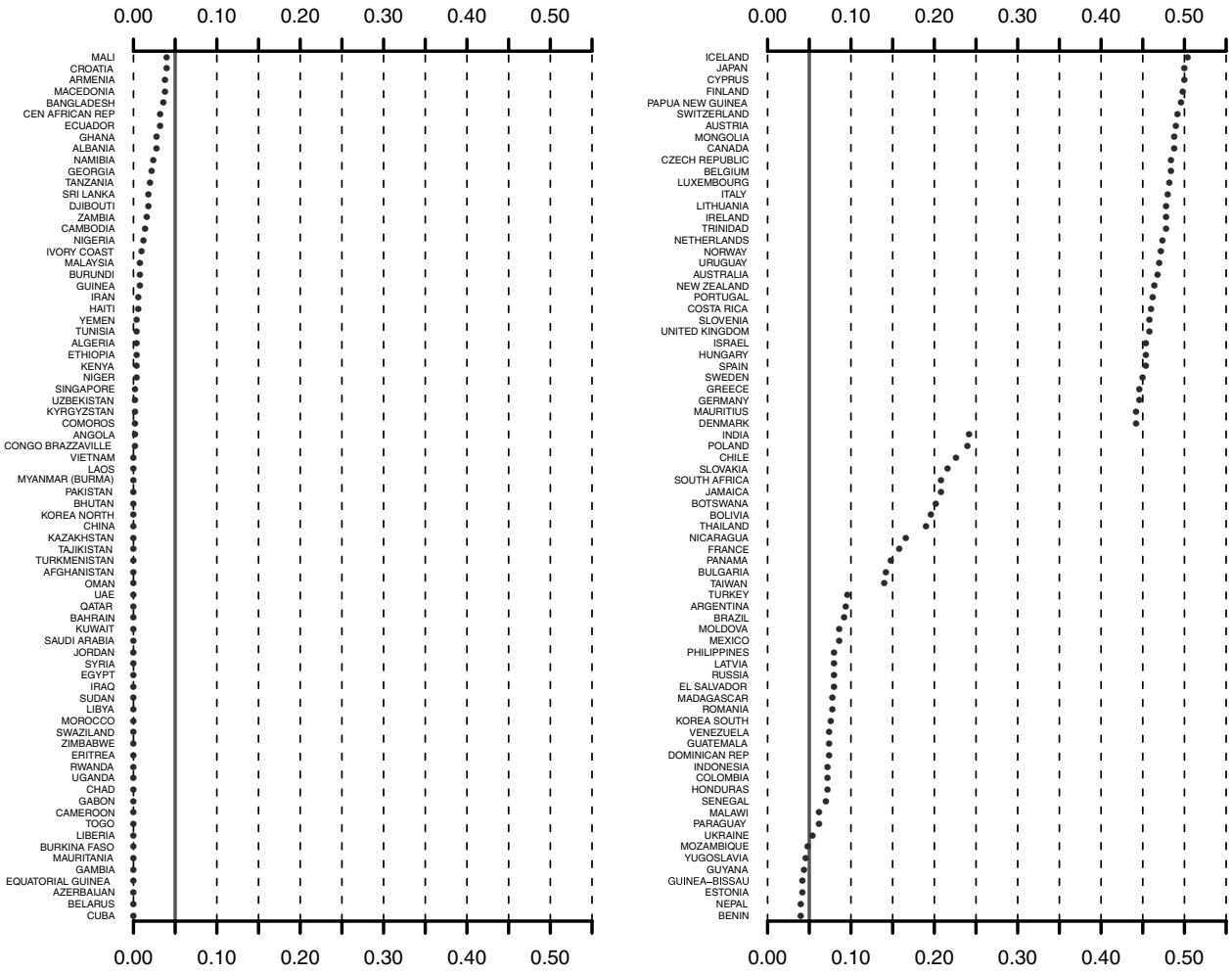
Figure 2 illustrates a considerable overlap in the HPD intervals for each country, suggesting that the uncertainty accompanying each estimate of the latent trait is large enough to make comparisons of latent levels of democracy difficult, in the sense that we cannot unambiguously make statements of the sort "country *a* has a higher level of democracy than country *b*." If that is the question (and it is a perfectly proper question to ask), then Figure 2 only tells part of the story. Since the latent traits are random variables, each with a marginal posterior density, the *difference* between any two latent traits x_i and x_j is also a random variable, with a variance equal to the variance

FIGURE 2 IRT Posterior Means for 2000



Notes: Countries are ordered by their posterior means. Error bars indicate 95% highest posterior density regions.

FIGURE 3 Probability of Higher Democracy Score than the United States, 2000



of x_i plus the variance of x_j minus twice the covariance between x_i and x_j. Figure 2 displays the pointwise confidence intervals of the latent levels of democracy, a function of the variances, but does not show anything about the covariances. To assess the precision with which we can make pairwise comparisons of levels of democracy, we compute the difference between the latent trait and that for the United States in the year 2000, i.e., $\delta_i = x_i - x_{\text{US}}$, and (of course) the uncertainty in that quantity.

In Figure 3 we graph the equivalent of a p-value for the one-sided hypothesis that $H_0 : \delta_i = x_i - x_{\text{US}} > 0$.[8] Seventy countries, or roughly one-half of the 153 countries available for analysis in 2000, have p-values greater than .05, implying that we cannot distinguish their democracy score from that for the United States at a conventional 95% level of statistical significance. Figure 3 reveals that there is a large cluster of countries which have democracy levels essentially indistinguishable from the United States in 2000; these include the advanced, industrial democracies of the OECD and other countries such as Mongolia, Costa Rica, Trinidad and Tobago, and Papua New Guinea. A second set of countries is more distinguishable from the United States, but we cannot determine at typical levels of significance that the United States is assuredly more democratic. Finally, there is little doubt that the remaining countries are less democratic than the United States.

[8]We compute this quantity by simply the proportion of times in repeated samples from the posterior density of the democracy measures we observe $\delta_i > 0$. Computing auxiliary quantities of interest such as these is remarkably simple in the Bayesian simulation approach we adopt.

The Consequences of Measurement Error: Democracy, Political Change, and Civil War

So we measure democracy imperfectly, with substantial amounts of measurement error. But how consequential is

this? That is, what inferential dangers are posed by using a measure of democracy in data analyses? As stated earlier, it is well known that using "noisy" variables in data analysis generates an "errors-in-variables" problem that will lead to biased and inconsistent parameter estimates, and potentially invalid hypothesis tests.

We explore the consequences of measuring democracy with error by replicating a recent study using the Polity data. Hegre et al. (2001) test hypotheses about the relationship between levels of democracy and civil war via duration analysis. Specifically, they use a Cox proportional hazards model to analyze the effect of democracy on time until a country experiences the outbreak of civil war. Their measure of democracy comes from the Polity IIId data collection, in which regime changes are recorded to the exact day whenever possible. A key hypothesis for Hegre et al. is that regimes with intermediate levels of democracy have a higher risk of outbreak of civil war than either democracies or autocracies. Earlier work has found evidence for this "U-shape" pattern between the occurrence or intensity of civil wars and various measures of democracy or repressiveness (e.g., de Nardo 1985; Ellingsen and Gleditsch 1997; Francisco 1995; Muller and Weede 1990). In contrast to democracies and autocracies, intermediate regimes are

> . . . partly open yet somewhat repressive, a combination that invites protest, rebellion, and other forms of civil violence. Repression leads to grievances that induce groups to take action, and openness allows for them to organize and engage in activities against the regime. Such institutional contradictions imply a level of political incoherence, which is linked to civil conflict. (Hegre et al. 2001, 33)

Likewise, intermediate regimes are often transitioning from autocracy to democracy, and regime change itself may be the destabilizing factor promoting civil war, rather than the intermediate level of democracy itself. Hegre et al. are careful to distinguish this possibility in their data analysis, including in their Cox regressions a control for temporal proximity to regime change. The main hypothesis—that intermediate regimes are at higher risk of civil war than either autocracies or democracies—is operationalized by including both the Polity score and its square in the Cox model. Other variables in the analysis include time since the country attained independence, a measure of ethnic heterogeneity, time since the country's last civil war, a measure of economic development (the log of energy consumption per capita, measured in coal-ton equivalents) and its square, and an indicator of whether the country was engaged in an interstate conflict (as defined in the Correlates of War Interstate War data set); see Hegre et al. (2001) for further details.

Our reanalysis is in two stages. First, using a data set made available by Hegre et al., we were able to exactly replicate their results as reported in their Table 2. Using the Polity measure of democracy, like Hegre et al., we found the coefficient on squared Polity score to be negative and distinguishable from zero at conventional levels of statistical significance ($p < .01$), while the coefficient on Polity score itself is swamped by its standard error.

We then reanalyzed the data using our measure of democracy, as follows. We first estimate our measurement model with the Polity IIId indicators, using the recoding scheme described earlier. Then, keeping all parts of the Hegre et al. analysis intact, we merged our measure of democracy into the Hegre et al. data set.[9] We then reestimated the Hegre et al. model using a Monte Carlo procedure to let uncertainty in levels of democracy propagate into inferences for the coefficients in the Cox regression model (see the appendix for details).

In Table 5 we report results for (1) the Hegre et al. model based on the 1946–92 data, exactly replicating their results; (2) the Hegre et al. model with the Syrian observations omitted; (3) replacing the Polity scores with the posterior means from the IRT analysis; and (4) allowing uncertainty as to a country's level of democracy propagate into inferences over the coefficients in the Hegre et al. Cox model. The consequences of acknowledging the uncertainty in a country's true level of democracy are quite dramatic in this instance: the coefficient on the square of democracy is no longer distinguishable from zero at conventional levels of statistical significance, while other coefficients in the model remain largely unchanged (i.e., comparing the parameter estimates in column 4 with the corresponding estimates in column 2 of Table 5). In short, one of the chief empirical findings of the Hegre et al. analysis is not replicated after we admit the uncertainty arising from measuring democracy with the Polity indicators.

Two distinct processes account for the way the Hegre et al. finding is not replicated with our measure. First, our reaggregation of the information in the Polity indicators creates more distinctions among countries than there are in the Polity scoring (by assigning different scores to

[9] In our analyses, we omit two regimes for Syria (1949–50 and 1954–58). The assigned Polity scores for these observations do not follow from the assigned values of the subindices. Since Hegre et al. use the Polity score and we use the subindices, we omit these observations in the analysis in order to avoid comparability problems. The omission of these cases leaves the results almost exactly unchanged.

TABLE 5 Risk of Civil War by Level of Democracy and Proximity of Regime Change, 1946–92

	Hegre et al. (2001)			
Explanatory Variable	**Original**	**Dropping Syria**	**IRT Point Estimates**	**Measurement Uncertainty Propagated**
Proximity of regime change	1.27	1.30	1.48	1.40
	(.47)	(.46)	(.46)	(.47)
Democracy	−.002	−.0025	−.187	−.029
	(.021)	(.021)	(.20)	(.17)
Democracy squared	−.012	−.012	−.095	−.11
	(.0051)	(.0051)	(.22)	(.15)
Proximity of civil war	1.16	1.10	1.21	1.20
	(.35)	(.35)	(.34)	(.35)
Proximity of independence	1.51	1.50	1.50	1.40
	(.97)	(.97)	(.96)	(.96)
International war in country	.86	.85	.93	.92
	(.59)	(.59)	(.51)	(.54)
Neighboring civil war	.097	.098	.16	.13
	(.33)	(.33)	(.32)	(.32)
Development	−.48	−.48	−.47	−.51
	(.15)	(.15)	(.16)	(.15)
Development squared	−.066	−.067	−.07	−.077
	(.035)	(.036)	(.037)	(.037)
Ethnic heterogeneity	.80	.78	.86	.91
	(.38)	(.38)	(.41)	(.40)

Notes: The first two columns report estimates of a Cox proportional hazards model, with robust standard errors in parentheses, clustered by country. Column 3 replaces the Polity scores with the posterior means of the estimates of latent democracy, without accounting for measurement uncertainty. Column 4 allows uncertainty in the democracy scores to propagate into inferences about the coefficients (posterior standard deviations are reported in parentheses).

every country with a distinct response pattern) and reorders the countries with respect to democracy. Second, in our approach, variability due to measurement uncertainty in the latent trait is dealt with explicitly, inducing additional variation in the point estimates from the duration analysis. To separate out the different effects of these two processes, we replace the Polity scores with just the posterior means from the IRT analysis, without propagating the measurement error into the duration analysis. Comparing these results (column 3, Table 5), it is apparent that the Hegre et al. findings are sensitive to our rescoring of democracy. That is, simply reaggregating the information in the Polity indicators—in a way implied by fitting a measurement model appropriate for these data—is sufficient to wash out the quadratic democracy term in the duration model. As shown in Figure 1, our model-based scoring procedure induces more separation between countries assigned the maximum and minimum Polity values. For countries assigned midrange Polity scores, our model-based procedure induces a more dispersed set of democracy scores. Conversely, recalling Table 4, many observations which were once separated by large distinctions on the Polity scale are much more similar according to the IRT scale. In short, the model-based scoring rule we use to aggregate the information in the Polity indicators produces a set of democracy scores that in turn stand in a different empirical relationship with an outcome like time until civil war onset, sufficient to generate estimated marginal effects that are indistinguishable from zero.[10]

The comparison between columns 3 and 4 in Table 5 highlights the biases that can result when ignoring measurement error. Allowing for the propagation of measurement error into the duration analysis substantially changes the magnitude of the coefficients for democracy. The coefficient on the linear democracy term in column 4 is swamped by its estimated standard error, and the coefficient on the quadratic term is smaller than its standard

[10]Nevertheless, even though the conclusions regarding statistical significance may not change, the magnitude of the marginal effects, which are the primary quantities of interpretative interest, will likely be different.

error (i.e., $|z| < 1$). The measurement error accompanying democracy is simply so large as to render it impossible to tease out a quadratic effect on democracy in the duration analysis. This should not be surprising: the sparse amount of information in the Polity indicators means that we can confidently resolve only large differences in democracy (e.g., Figure 3), so little wonder that we fail to be able to resolve a quadratic relationship on democracy in the duration analysis.

The lesson here is that taking Polity scores at face value is tantamount to pretending that we know more than we do about democracy, and, at least in this case, a structured approach to the measurement of democracy leads to a measure that is sufficiently different from Polity and sufficiently "noisy" to disrupt one finding in the literature. This said, we stress that our results do not falsify the U-curve theory and are certainly not the last word in the debate about the relationship between democracy and civil war onset.[11] What the results do indicate is that one's conclusions can change dramatically if we do not properly account for error in our measurements, and researchers must consider the possibility that their conclusions depend on the quality of their operationalizations.

We also stress that this result—the propagation of uncertainty as to underlying levels of democracy leading to a statistically insignificant estimate of the effect of democracy on a dependent variable—is somewhat rare. In other replication experiments we have noted that a proper accounting of measurement uncertainty leads to a diminuation of the effect one would associate with democracy on a particular outcome, but not enough to render the estimated effect of democracy indistinguishable from zero. This is because in many applications (1) the estimated effects of democracy are very strong and can withstand any attenuation or increased parameter uncertainty due to the propagation of uncertainty arising from the imperfect measurement of democracy; and (2) the statistical models being deployed are relatively simple (e.g., a country's Polity score enters as a single linear term). Our experience is that the risks of "pretending we know more about levels of democracy than we really do" bite when researchers rely on elaborations such as nonlinear functional forms (e.g., the Hegre et al. analysis relies on democracy entering the duration analysis via a quadratic) or highly interactive specifications. In these cases, it is not surprising that the purported effects of democracy dissolve in the face of measurement uncertainty: given the uncertainty that accompanies extant measures of democracy, we simply will not be able to resolve a relatively flamboyant functional form on democracy in a regression-type analysis. In short, there simply is not enough information in the Polity indicators to support particularly elaborate models of the way democracy structures outcomes.

[11] Indeed, Vreeland (2005) even argues that the collection of Polity makes the indicators inappropriate for studying civil wars. He demonstrates the coding of Polity is endogenous to the dependent variable; countries in civil wars by definition are coded as transitional regimes, which are assigned particular values (in the middle) on the Polity indicators. Thus, countries suffering an upheaval will be automatically classified as semi-democracies.

Conclusion

Even though the Polity data have been used in hundreds of studies of comparative politics and international relations, some scholars are skeptical of the properties of the measure, and rightly so. Using a formal, statistical measurement model, we show how to make best use of the Polity indicators, leveraging their strengths against one another, to obtain estimates of a given country's underlying level of democracy. Our approach—an ordinal item-response model—improves upon the widely used Polity democracy scale in several respects. Like a factor analytic approach, we rely on the relationships among the Polity indicators to tell us how to weight each indicator's contribution to the score we assign for any given country; our item-discrimination parameters are the equivalent of factor analysis' factor loadings. But unlike conventional factor analytic models, we embed each country's level of democracy as an unknown parameter in the measurement model, and recover not only point estimates, but also the entire joint distribution of democracy scores for all countries. Assessments of measurement error and its consequences are easily obtained via this approach. We show that there is considerable error in the latent levels of democracy underlying the Polity scores. Moreover, this measurement error is heteroskedastic; countries found to have extremely high or low levels of democracy also have the most noisy measures of democracy. The consequences are that when we use democracy as an independent variable, but ignore the noise in the democracy measure, the risk of inferential error is high. For instance, in replicating a simple duration analysis relating the level of democracy and the outbreak of civil war, we find that an apparently quadratic relationship is not robust after we properly account for the measurement error in the democracy variable.

We close with two recommendations. First, it is apparent that we need more and/or better indicators of democracy. In this analysis, we rely on five indicators in the Polity data set, effectively reduced to three indicators due

to inherent dependencies in the coding of the indicators. Accordingly, we are measuring democracy with a fairly blunt set of tools; contrast other measurement exercises in political science, say survey-based measures of ideology formed from aggregating 10 to 20 self-placement items (each with 7-point scales), or recovering estimates of legislative preferences from roll-call data (e.g., each session of the U.S. Congress yields hundreds or even thousands of roll calls, giving us considerable ability to distinguish legislators from one another). Consequently, in our application, any possible effect of democracy is subsumed by the overwhelming amount of uncertainty present in the democracy scores. Adding even a few more indicators could improve the reliability of democracy measures considerably. More indicators would imply greater distinctions between observations, and reduce the amount of uncertainty associated with the scores. And by utilizing an appropriate statistical model, aggregating scores involves no additional complications, unlike the problems that occur when creating additive indices.

One could also complement this strategy by moving to a multiple rater system, asking area specialists to give scores on the various indicators (including the existing Polity indicators). Ward (2002) and Bollen and Paxton (2000) illustrate the variability in subjective judgments by coders, as well as potential biases that can arise. Thus, relying on the subjective judgments of one coder can be problematic. A design incorporating multiple raters would have the virtue of not only letting us leverage the indicators against one another (as we do now), but would also let us leverage expert opinions against one another.[12] This would be one way of expanding the amount of data available for measuring democracy.

Second, while a better measure of democracy is a scientific advance in and of itself, it is even more important to consider the consequences of working with a necessarily imperfect measure of democracy. The methodology we present in this article provides a simple recipe for avoiding the overoptimism that can result when working with noisy measures.[13] Failing to properly acknowledge the measurement uncertainty in latent constructs risks inferential errors; scholars finding significant impacts of democracy on various dependent variables may well be wrong or (at least) guilty of overstating matters, pretending that they know more about a country's level of democracy than they really do. Whatever measure of democracy one uses, and however one derives it, we strongly recommend using methods like those we deploy here, ensuring that inferences about the effect of democracy on an outcome variable reflect the fact that a country's level of democracy is the product of an imperfect measurement process, and hence uncertain and error-prone. Like so many concepts in social science, a country's level of democracy is a fiction of sorts, a manufactured construct, an abstraction rendered in a form amenable for data analysis: the tools we present here let us stop pretending otherwise.

[12] See Jackman (2004) and Clinton and Lewis (n.d.) for applications of multiple rater models.

[13] The method also extends to cases where there is more than one latent variable; see Lee (2007) for a fully Bayesian implementation of these structural models, including situations when the data are discrete or not normally distributed.

Appendix

Identification and Estimation

Since the likelihood for the ordinal item-response model is parameterized in terms of the combination of latent constructs and item parameters $\mu_{ij} = x_i\beta_j$, changes in the x_i can be offset by changes in the β_j, yet provide the same likelihood contributions. In particular, $\mu_{ij} = x_i\beta_j = (x_i r)\ (\beta_j r^{-1})$ for any $r \neq 0$. Further, the latent levels of democracy x_i can all be shifted by some constant c, yielding $\mu_{ij} = x_i\beta_j + c\beta_j$, with offsetting shifts in the threshold parameters $\boldsymbol{\tau}_j$ yielding the same likelihood. To solve this lack of identification, we constrain the latent x_i to have mean zero and variance one, ruling out arbitrary shifts in location and scale for the latent traits, providing local identification, following the definition in Rothenberg (1971).

In the Bayesian approach, which simplifies the estimation and inference for this model, interest centers on the joint posterior density of the model parameters, $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$. Subject to regularity conditions, Markov-chain Monte Carlo methods generate a random tour of the parameter space, visiting regions with frequency proportional to their posterior probability. Thus, summaries of the trajectory of a long, MCMC-generated random tour amount to summaries of the joint posterior density. Estimation and inference is straightforward: we compute point estimates of latent levels of democracy by simply averaging the output of many iterations of the MCMC algorithm for the x_i parameters. Assessments of the magnitude of the measurement error are obtained by computing the dispersion of the posterior density of each x_i parameter by calculating the standard deviation of the MCMC output with respect to the x_i parameters.

The MCMC random tour of the parameter space is generated by successively sampling from the conditional distributions that together characterize the joint

posterior density. This is helpful since the constituent conditional distributions are of much lower dimension than the joint posterior density. For our ordinal IRT model, iteration t of the MCMC algorithm involves sampling from the following three sets of conditional distributions:

1. sample $x_i^{(t)}$ from $g_x(x_i \mid \boldsymbol{\beta}^{(t-1)}, \boldsymbol{T}^{(t-1)}, \boldsymbol{Y})$, $i = 1, \ldots, n$
2. sample $\beta_j^{(t)}$ from $g_\beta(\beta_j \mid \boldsymbol{x}^{(t)}, \boldsymbol{\tau}_j^{(t-1)}, \boldsymbol{Y})$, $j = 1, \ldots, m$
3. sample $\boldsymbol{\tau}_j^{(t)}$ from $g_{\boldsymbol{\tau}}(\boldsymbol{\tau}_j \mid \boldsymbol{x}^{(t)}, \beta_j^{(t)}, \boldsymbol{Y})$, $j = 1, \ldots, m$.

MCMC algorithms for ordinal response models are described in greater detail in Johnson and Albert (1999, 133–36). We implement this MCMC scheme using `WinBUGS` (Lunn et al., 2000).

Priors. We employ normal priors for the discrimination parameters β_j with mean zero and variance 9. Our priors for the threshold parameters $\mathbf{T}$ are also chosen to reflect prior ignorance. Ordering constraints are imposed by parameterizing the thresholds as $\tau_{jk} = \sum_{l=1}^{k} \delta_{jl}$, $k = 1, \ldots, K_j$, and $\delta_{jk} > 0$, where $k \geq 2$. The first cutpoint, $\tau_{j1} \equiv \delta_{j1}$, follows a normal prior with mean zero, variance $6\frac{2}{3}$. The subsequent quantities δ_{jk}, $k \geq 2$, are assigned exponential priors with mean $\frac{1}{2}$. We specify $N(0, 1)$ priors on each x_i, but after updating the x_i at each iteration, center and scale the x_i to have mean zero and variance one. Operationally, we impose the recentering and rescaling of the x_i on the output, iteration-by-iteration, effectively "post-processing" the MCMC output (e.g., Hoff, Raftery, and Handcock 2002; de Jong, Wiering, and Drugan 2003). Of course, transforming the x_i this way implies that the β_j and $\mathbf{T}$ be appropriately transformed. We initialize the MCMC algorithm with start values for the country-year using the original Polity score divided by ten; for β_j and the thresholds, we use the estimates of the cutpoints and slope parameters from an ordered logit model of the Polity indicators on the start values for country-year. After discarding the first 10,000 iterations as burn-in, estimates and inferences are based on 50,000 iterations, thinned by 100, in order to produce 500 approximately independent draws from the joint posterior density.

Hegre et al. reanalysis: Our Bayesian measurement procedure yields the posterior density of x_{it}, levels of democracy in country i in year t. Let $\mathbf{x}$ be a vector containing the x_{it}. Denote the posterior density of $\mathbf{x}$ as $p(\boldsymbol{x} \mid \boldsymbol{Z})$, where $\mathbf{Z}$ are the Polity indicators. Let $\mathbf{y}$ denote the durations to be modeled and $\boldsymbol{\beta}$ denote the parameters in the duration model. The predictors in the duration model are of two types: $\mathbf{x}$ (levels of democracy) and other controls, which we denote as $\mathbf{w}$. If $\mathbf{x}$ were known without measurement error, then inference for $\boldsymbol{\beta}$ is unproblematic; one could simply estimate the Hegre et al. Cox model in the conventional way (i.e., via partial likelihood methods), and, following Hegre et al., use an asymptotically consistent estimator of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, clustering on the countries in the analysis. However, levels of democracy are not measured perfectly, and known only up to a distribution, in the posterior distribution $p(\mathbf{x} \mid \mathbf{Z})$.

For measurement uncertainty in $\mathbf{x}$ to propagate into inferences over the parameters in the duration model, we employ the following iterative Monte Carlo procedure: at iteration t

1. Sample $\mathbf{x}^{(t)}$ from the posterior distribution $p(\mathbf{x} \mid \mathbf{Z})$.
2. Run the Hegre et al. duration model, with durations $\mathbf{y}$ and predictors $\mathbf{x}^{(t)}$ and $\mathbf{w}$. This yields partial likelihood estimates of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}^{(t)}$, and the estimated "robust" variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{(t)}$, $\hat{\mathbf{V}}^{(t)}$. As $\mathbf{x}^{(t)}$ changes over iterations, reflecting measurement uncertainty in $\mathbf{x}$, so too will $\hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\mathbf{V}}^{(t)}$.
3. Sample $\tilde{\boldsymbol{\beta}}^{(t)}$ from the multivariate normal density with mean vector $\hat{\boldsymbol{\beta}}^{(t)}$ and variance-covariance matrix $\hat{\mathbf{V}}^{(t)}$.

Each iteration yields $\tilde{\boldsymbol{\beta}}^{(t)}$, a sample from a density which can be considered the posterior density for $\boldsymbol{\beta}$, taking into account both measurement uncertainty in $\mathbf{x}$ (levels of democracy) and uncertainty about the effects of the predictors $\mathbf{x}$ and $\mathbf{w}$ on the durations $\mathbf{y}$. More formally, the algorithm provides a way to sample $\boldsymbol{\beta}$ and $\mathbf{x}$ from their joint posterior density via the decomposition

$$p(\boldsymbol{\beta}, \mathbf{x} \mid \mathbf{w}, \mathbf{y}, \mathbf{Z}) = p(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{w}, \mathbf{y})\, p(\mathbf{x} \mid \mathbf{Z}),$$

i.e., writing a joint density over $(\boldsymbol{\beta}, \mathbf{x})$ as a product of a conditional density and a marginal density. Step 1 provides samples from the second density on the right-hand side of this equation, while step 3 provides samples from the first density, the marginal posterior density for $\boldsymbol{\beta}$, i.e.,

$$p(\boldsymbol{\beta} \mid \mathbf{w}, \mathbf{y}) = \int_{\mathcal{X}} p(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{w}, \mathbf{y})\, p(\mathbf{x} \mid \mathbf{Z})\, d\mathbf{x}$$

where the iterative algorithm performs the intergration by the Monte Carlo method. See Tanner (1996, 52) for further details on this technique, known as the "method of composition." Note also the implicit assumptions at

work here: (1) $p(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{w}, \mathbf{y}) = p(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{w}, \mathbf{y}, \mathbf{Z})$ (i.e., the Polity indicators $\mathbf{Z}$ do not supply information about $\boldsymbol{\beta}$ directly, but only through levels of democracy, $\mathbf{x}$); (2) $p(\mathbf{x} \mid \mathbf{Z}) = p(\mathbf{x} \mid \mathbf{w}, \mathbf{y}, \mathbf{Z})$ (i.e., the durations $\mathbf{y}$ and other predictors $\mathbf{w}$ do not supply information about levels of democracy $\mathbf{x}$ beyond that in the Polity indicators $\mathbf{Z}$). Assumption (1) is reasonable; assumption (2) seems slightly less plausible, in that it separates *measurement* of democracy from the use of democracy in subsequent modeling, whereas if one believes durations are a function of democracy, and democracy is measured imperfectly, then durations are informative with respect to democracy. Here we have adopted the more restrictive assumption (2), focusing on the quality of the Polity indicators as measures of democracy.

References

Bernardo, José, and Adrian F. M. Smith. 1994. *Bayesian Theory*. 1st ed. Chichester: Wiley.

Bollen, Kenneth A. 1993. "Subjective Measures of Liberal Democracy." *American Journal of Political Science* 33(1): 58–86.

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

Carroll, R. J., D. Ruppert, and L. A. Stefanski. 1995. *Measurement Error in Nonlinear Models*. 1st ed. New York: Chapman and Hall.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2): 355–70.

Clinton, Joshua D., and David E. Lewis. n.d. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis*. Forthcoming.

Coppedge, Michael, and Wolfgang H. Reinicke. 1991. "Measuring Polyarchy." *Studies in Comparative International Development* 25(1): 51–72.

de Jong, Edwin D., Marco A. Wiering, and Mădălina M. Drugan. 2003. Post-Processing for MCMC. Technical Report UU-CS-2003-021, Institute of Information and Computing Sciences, Utrecht University.

de Nardo, James. 1985. *Power in Numbers*. Princeton, NJ: Princeton University Press.

Ellingsen, Tanja, and Nils Petter Gleditsch. 1997. "Democracy and Armed Conflict in the Third World." In *Causes of Conflict in Third World Countries*, ed. Ketil Volden and Dan Smith. Oslo: North-South Coalition and International Peace Research Institute, 69–81.

Erikson, Robert S. 1990. "Roll Calls, Reputations, and Representation in the U.S. Senate." *Legislative Studies Quarterly* 15 (4): 623–42.

Francisco, Ronald A. 1995. "The Relationship between Coercion and Protest: An Empirical Evaluation in Three Coercive States." *Journal of Conflict Resolution* 39(2): 263–82.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York: John Wiley and Sons.

Gasiorowski, Mark J. 1996. "An Overview of the Political Regime Change Dataset." *Comparative Political Studies* 29(4): 469–83.

Gleditsch, Kristian S., and Michael D. Ward. 1997. "Double Take: A Reexamination of Democracy and Autocracy in Modern Polities." *Journal of Conflict Resolution* 41(3): 361–83.

Green, Donald P., Alan S. Gerber, and Suzanna L. De Boef. 1999. "Tracking Opinion Over Time: A Method for Reducing Sampling Error." *Public Opinion Quarterly* 63(2). 178–92.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

Hegre, Håvard, Tanja Ellingsen, Scott Gates, and Nils Petter Gleditsch. 2001. "Toward a Democratic Civil Peace? Democracy, Political Change, and Civil War, 1816–1992." *American Political Science Review* 95(1): 33–48.

Hoff, Peter, Adrian E. Raftery, and Mark S. Handcock. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97(460): 1090–98.

Huber, John, and Ronald Inglehart. 1995. "Expert Interpretations of Party Space and Party Locations in 42 Societies." *Party Politics* 1(1): 73–111.

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(4): 400–424.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.

Lee, Sik-Yum. 2007. *Structural Equation Modelling: A Bayesian Approach*. Nw York: John Wiley & Sons.

Lunn, David J., Andrew Thomas, Nicky Best, and David J. Spiegelhalter. 2000. "WinBUGS—a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10(4): 325–37.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. San Diego: Academic Press.

Marshall, Monty G., Ted Robert Gurr, Christian Davenport, and Keith Jaggers. 2002. "Polity IV, 1800–1999: Comments on Munck and Verkuilin." *Comparative Political Studies* 35(1): 40–45.

Marshall, Monty G., and Keith Jaggers. 2002a. "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2000. Dataset Users Manual." Retrieved from http://www.cidcm.umd.edu/inscr/polity/.

Marshall, Monty G., and Keith Jaggers. 2002b. "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2000. The Polity IV Dataset." Retrieved from http://www.cidcm.umd.edu/inscr/polity/.

Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2): 134–53.

Muller, Edward N., and Erich Weede. 1990. "Cross-National Variations in Political Violence: A Rational Action Approach." *Journal of Conflict Resolution* 34(4): 624–51.

Munck, Gerardo L., and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35(1): 5–34.

Reckase, Mark D. 1997. "The Past and Future of Multidimensional Item Response Theory." *Applied Psychological Measurement* 21(1): 25–36.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39(3): 577–91.

Takane, Yoshio, and Jan de Leeuw. 1987. "On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables." *Psychometrika* 52(3): 393–408.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 3rd ed. New York: Springer-Verlag.

Vreeland, James R. 2005. "Research Note: A Problem with Polity—Unpacking Anocracy." Typescript. Yale University.

Wansbeek, Tom, and Erik Meijer. 2000. *Measurement Error and Latent Variables in Econometrics*. Amsterdam: North-Holland.

Ward, Michael D. 2002. "Green Binders in Cyberspace: A Modest Proposal." *Comparative Political Studies* 35(1): 46–51.