

Democracy as a Latent Variable*

Shawn Treier
Stanford University
satreier@stanford.edu

Simon Jackman
Stanford University
jackman@stanford.edu

July 17, 2003

Abstract

Measurement is critical to the social scientific enterprise. Many key concepts in social-scientific theories are not observed directly, and researchers rely on assumptions (tacitly or explicitly, via formal measurement models) to operationalize these concepts in empirical work. In this paper we apply formal, statistical measurement models to the Polity IV data, a set of country-level indicators of democracy. In so doing, we make explicit the hitherto implicit assumptions underlying scales built using the Polity indicators. We apply two models: one in which democracy is operationalized a latent continuous variable, and another in which democracy is operationalized a latent class. We show how to better exploit the information in the Polity data set so as to produce a more reliable scale measure (or classification) of democracy. Our modeling approaches also let us assess the “noise” (measurement error) in our resulting measure of democracy. 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 analysis. 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.

1 Latent Variables Abound in Political Science

Social and political theories often refer to constructs that can not be observed directly. Examples include public opinion, socio-economic status, social capital, ideology, or democracy. Instead of observing these quantities, researchers may have *indicators* of these concepts,

*Prepared for delivery at the 2003 Annual Meeting of the Society for Political Methodology, University of Minnesota, Minneapolis, July 17-19, 2003. Earlier versions of this work were presented at the 2003 Annual Meeting of Midwestern Political Science Association and at Stanford University. We thank Jon Bendor, Alberto Diaz, Jim Fearon, Steve Krasner, David Laitin, Andrew Martin, Doug Rivers, and Mike Tomz for useful comments and references. Errors and omissions remain our own responsibility.

tapping the latent concept with varying degrees of fidelity. Negotiating the leap from the observed social world to theoretically interesting concepts is part and parcel of social scientific practice. Theoretical concepts are often rich and multi-faceted, whose substantive content is seldom totally captured by the available indicators: e.g., there is more to religiosity than frequency of church attendance, there is more to socio-economic status than income, there is more to state power than the size of a country's armed forces. The standard approach to these measurement problems is to use statistical procedures to combine the information in multiple indicators of the latent concept; indeed, formalizing procedures for combining information from multiple sources is one of the most valuable achievements of twentieth century social science.

Political scientists combine information from multiple indicators in several ways. Researchers often create an linear additive scale (or some other weighting scheme fixed *a priori* by the researcher), simply summing each indicator, perhaps weighting or re-scaling each item so that the contributions of each item to the scale are equal. A prominent example -- and the subject of this paper -- is in constructing measures of the level of democracy in countries. Measures of democracy are used extensively in empirical work on the "democratic peace" and economic development. 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 extant measures of democracy. Scientific measurement begins by asking how to operationalize the latent construct, and so we survey different conceptualizations of democracy, showing how each leads to different statistical measurement models, one in which democracy is a latent continuous variable, and one in which democracy is a latent binary variable. We contrast the measures of democracy provided by these formal measurement models with extant measures. We quantify uncertainty in our resulting measures of democracy (which is considerable), and demonstrate the consequences of measurement error for our ability to estimate the impact of democracy on outcomes of substantive interest such as interstate conflict.

2 Democracy: concepts and measures

A tremendous amount of time and effort has been devoted to measuring democracy, or 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?

Predictably, the theoretical literature on democracy does not speak with one voice on this question. In the introduction to *A Preface to Democratic Theory*, Dahl explicitly avoids a precise definition of democracy "... for each of the chapters is to some extent an essay in definition" (p2), but throughout Dahl's writing one comes across the notion that democracy is something that can be attained to greater or less extent. Having eschewed a rigorous definition of democracy, Dahl goes on to state that "at a minimum, ..., democratic theory is concerned with processes by which ordinary citizens exert a *relatively high degree* of control over leaders" (Dahl, 1956, 3, emphasis added). Elsewhere, Dahl explicitly refers to ordering countries along a theoretical scale of democracy according to the presence or absence of various institutional guarantees (i.e., indicators), noting that

both historically and contemporaneously, regimes also vary in the proportion of the population entitled to participate.... A scale reflecting the breadth of the right to participate in public contestation would enable us to compare different regimes according to their inclusiveness (Dahl, 1971, 4).

Dahl refers to the most inclusive of real-world political systems as "polyarchies", since "no large system in the real world is fully democratized" Dahl (1971, 8), again suggesting that for Dahl, democracy is a latent, continuous quantity with an (as yet unrealized) upper bound. Almost regimes exhibit some "degree of democratization" and polyarchies may be thought of as "relatively (but incompletely) democratized regimes" (Dahl, 1971, 8). Further, Dahl is reluctant to classify regimes as "hegemonic" or "polyarchic" since "the arbitrariness of the boundaries between 'full' and 'near' testifies to the inadequacy of any classification" (Dahl, 1971, 9).

Nonetheless, many scholars conceptualize democracy as a dichotomy. Examples of this approach in political science include Lipset (1960), Linz (1975), Powell (1982) and Huntington (1991). According to Powell (1982, 4), "despite important variations in the degree of democracy, the distinction between contemporary nations meeting most of the ...[democratic criteria]... most of the time and those failing to do so is fairly clear." A large literature in sociology examining cross-national differences in income inequality sees authors such as Rustow (1967), Hewitt (1977) and Muller (1988) treat democracy as a dichotomy, while Bollen (1980) and Bollen and Jackman (1989) argue forcefully that democracy is a latent continuous variable:

Dichotomizing democracy lumps together countries with very different degrees of democracy and blurs distinctions between borderline cases. For example, are democratic practices entirely absent from Mexican politics? ... The difficulty

in answering these questions reflects the inherently continuous nature of the concept of political democracy (1989, 612).

The tension between dichotomous and continuous operationalizations of democracy was revived in political science after a forceful intervention by Alvarez et al. (1996). According to ALCP, democracy is like the “proverbial pregnancy” in that “while democracy can be more or less advanced, one cannot be half-democratic: there is a natural zero point” (Alvarez et al., 1996, 21). Thus, proponents of the view that democracy is a latent continuous variable such as Bollen and Jackman are “confused” (Alvarez et al., 1996, 21), apparently not sharing the ALCP view that democracy is “first a question of kind before it is one of degree” (to use Elkins’ (2000) paraphrase of the ALCP position). For ALCP, the dispositive requirement for democracy is that incumbents lose elections (i.e., there are competitive elections). In their applied work, ALCP contrast the performance of democracies with dictatorships on development outcomes, unburdened by any shades of gray or borderline cases.

In this paper we need not pick sides in the “continuous-versus-dichotomy” controversy. Our contribution here is to show that whatever conceptualization one might choose, there are principled, statistical methods for using *indicators* of democracy to arrive at either measures or classifications of regimes. Our position is that democracy is a latent variable, and can not be measured directly, but that indicators of democracy (of varying degrees of fidelity) are available. The statistical issues we address here are (1) how to best exploit the information in the indicators so as to arrive at a summary measure of the democracy in a regime (either on a continuum or classification into a binary state); (2) how to ensure that whatever uncertainty exists in the resulting measure or classification¹ of democracy propagates into subsequent statistical uses of the measure. In particular, we show how to guard against developing a false sense of security when using measures of democracy as, say, an independent variable in a regression analysis; inferences about the effects of democracy on some outcome of interest ought to reflect the fact that democracy is a latent variable, measured indirectly via a finite number of imperfect indicators and hence subject to measurement error.

2.1 Other Issues

Of course, the distinction between dichotomous and continuous notions of democracy does not exhaust the conceptual or operational possibilities. Some researchers operationalize democracy with just a single indicators, seeing approaches based on multiple indicators as unnecessarily complicated (Gasiorowski, 1996, e.g.). Among scholars who operationalize

¹Unlike Alvarez et al. (1996), our classification of regimes is probabilistic/statistical rather than deterministic (based on application of ALCP’s coding rules), and hence subject to uncertainty.

democracy via multiple indicators, there is considerable variation in how to *aggregate* or perform *data reduction*, the task of assigning a single score to each country-year, given the scores on the indicators. Perhaps the most widely used set of democracy scores, Polity [Marshall and Jagers \(2002b\)](#), combines indicators in a linear, additive way. Other scholars have used more sophisticated methods of aggregations, such as Guttman scaling or other statistical methods of data reduction such as factor analysis (e.g., [Bollen, 1980](#); [Bollen and Paxton, 2000](#)). This considerable variation in measurement procedures suggests that there seems to be no settled method of aggregating indicators of democracy. A second and related problem is that scholars who either create or use measures of democracy seldom confront the issue of measurement error. Whenever democracy appears as an explanatory variable in empirical work, there is an almost-always-ignored errors-in-variables problem, potentially invalidating the substantive conclusions of these studies. The approach we present below explicitly confronts the fact that like any latent variable, democracy is measured with error; we show how some recent studies of international conflict warrant reassessment in light of the measurement error inherent in democracy.

Finally, another issue in operationalizing democracy is the question of dimensionality. We do not explore this issue in detail in this paper, but note that some theoretical accounts, such as [Dahl \(1971\)](#), are explicitly multidimensional. While numerous empirical studies have adopted a multidimensional approach (CITES), the bulk of the empirical literature on democracy operationalizes it as a unidimensional construct. We follow that approach here, and offer some empirical evidence to suggest that this reasonable.

3 The Polity IV 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 Polity collection from the [Polity IV Project \(Marshall and Jagers, 2002b\)](#). These scores are used extensively in international relations and comparative politics; more importantly, all of the indicators used to construct the aggregate measure are accessible and well documented, unlike some other alternative measures. The Polity scores are designed to provide a scaled description of “polities,” based on “authority patterns” based on the theoretical work by [Eckstein and Gurr \(1975\)](#). The observed data are indicators related to executive recruitment, directiveness and responsiveness, constraints on the executive, and political participation. The Polity IV scores

use five expert-coded categorical indicators, all capable of being ordered: they are²

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).

The Polity IV indicators cover the period of 1800--2000 for some 184 countries, for a total of 13,198 country-years. The summary measure used widely in empirical applications is a country-year's "Polity score", and is computed as the difference of two sub-indices, "Democracy" and "Autocracy", which are created using the addition rules presented in Table 1. Both sub-indices range between 0 and 10 and are integer valued. The Polity measure is calculated as the difference between the Democracy and the Autocracy indices, and ranges from -10 to 10, inclusive.

The Polity measure has some oddities warranting elaboration. Specific categories on the indicators contribute pre-assigned scores to the Democracy and Autocracy sub-indices. But not all categories on every indicator makes a contribution to either sub-index, leading to some inconsistencies in the final Polity score. For instance, the category "Unregulated" for both XROPEN and XRCOMP contributes zero to both the both the Democracy and Autocracy sub-indices. By coding these categories this way, and constructing the Polity score as the difference between Democracy and Autocracy sub-indices, the "Unregulated" category is implicitly being counted as an intermediate category for XROPEN and XRCOMP in their contribution to the Polity score; however, the description of these categories in [Marshall and Jagers \(2002a\)](#) clearly indicates that the "Unregulated" category should clearly be the bottom category, and we treat it as such.³

²Additional information on the dataset and the coding of the variables is available in [Marshall and Jagers \(2002a\)](#)

³The "Unregulated" categorization results from a sixth indicator, Regulation of Chief Executive Recruitment (XRREG), which is an ordered variable with three categories, acts as a filter variable for XRCOMP and XROPEN. If XRREG indicates that the polity-year observation is not "regulated" (top category), but coded either as "Unregulated" (bottom category) or "Designational/Transitional" (middle category), then both XRCOMP and XROPEN are coded as zero. "Unregulated" refers to regimes in which changes in the chief executive are through forceful seizures of power; "Designational/Transitional" refers to noncompetitive designation, forceful seizures followed by transitional arrangements, or cases where there is a seizure of power followed shortly by rigged elections. It is clear from the description of the indicators that these cases are closer to the Autocracy ideal-type than the other codings; therefore, the "Unregulated" category should be the bottom case. In all of our analyses, we treat it as such.

	Contribution to Democracy	Contribution to Autocracy	Contribution to Polity	Our Ordering
Competitiveness of political participation (PARCOMP):				
Competitive	3	0	3	5
Transitional	2	0	2	4
Factional	1	0	1	3
Restricted	0	1	-1	2
Suppressed	0	2	-2	1
Not applicable	0	0	0	NA
Regulation of political participation (PARREG):				
Regulated	0	0	0	4
Multiple Identity	0	0	0	3
Sectarian	0	1	-1	2
Restricted	0	2	-2	1
Unregulated	0	0	0	NA
Competitiveness of executive recruitment (XRCOMP):				
Election	2	0	2	4
Transitional	1	0	1	3
Selection	0	2	-2	2
Unregulated	0	0	0	1
Openness of executive recruitment (XROPEN):				
Election	1	0	1	5
Dual: Hereditary and Election	1	0	1	4
Dual: Hereditary and designation	0	1	-1	3
Closed	0	1	-1	2
Unregulated	0	0	0	1
Constraints on Chief Executive (XCONST):				
Parity or Subordination	4	0	4	7
Intermediate 1	3	0	3	6
Substantial	2	0	2	5
Intermediate 2	1	0	1	4
Slight moderation	0	1	-1	3
Intermediate 3	0	2	-2	2
Unlimited Power	0	3	-3	1

Table 1: Description of Polity Coding Rules. Adapted from Table 1 in [Marshall et al. \(2002\)](#).

There are two other coding issues. First, PARCOMP has a “Not Applicable” category, and these observations are coded as “Unregulated” in the PARREG variable. In this case, “Unregulated” implies “fluid” participation. It does not imply there is no participation, just that it is unstructured and idiosyncratic. It is not clear whether this case is less democratic than the bottom category for PARREG, “Repressed,” or even if the category actually fits in the ordered categories at all. Thus, we treat cases as missing data on these two indicators, and impute as MAR data. Second, there are three missing value codes, indicative of periods of interruption, interregnum, or transition. In any year where these periods occur, all of the indicators are missing. Consequently, we remove these observations from the analysis.

These oddities aside, the more fundamental deficiency of Polity is the arbitrariness of the aggregation or data reduction rule. For the most part, the contribution of each category for each indicator increments linearly; it is usually the case that any one category increase/decrease on any one of the ordinal indicators yields a unit increase/decrease in the Polity score, save for the two or three unit jumps in the middle of some indicators (e.g., examine the respective contributions of PARCOMP, XRCOMP, XROPEN, and XCONST to the Polity score in Table 1). But is this the most appropriate aggregation rule for these indicators? Where are the principled, theoretical justifications for this particular scoring procedure? Can the Polity indicators be treated as interval measures? Should moving from a 1 to 2 on indicator j have the same contribution to the resulting measure of democracy as, say, moving from, 3 to 4 on indicator k ? Moreover, do all indicators tap the latent construct (democracy) equally well? In short, to what extent is the data reduction rule employed by Polity supported by the data? In the empirical analysis which follows, we present a statistical model that lets the data themselves speak to these issues.

4 Ordinal Item-Response Model

Our empirical analysis begins by treating 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 an ordinal item-response model. Let $i = 1, \dots, n$ index country-years and $j = 1, \dots, m = 5$ index the Polity indicators. Let $k = 1, \dots, K_j$ index the (ordered)

response categories for item j . Then our model is

$$\begin{aligned}
 \Pr(y_{ij} = 1) &= F(\tau_{j1} - \mu_{ij}) \\
 \Pr(y_{ij} = 2) &= F(\tau_{j2} - \mu_{ij}) - F(\tau_{j1} - \mu_{ij}) \\
 &\vdots \\
 \Pr(y_{ij} = k) &= F(\tau_{jk} - \mu_{ij}) - F(\tau_{j,k-1} - \mu_{ij}) \\
 &\vdots \\
 \Pr(y_{ij} = K_j) &= 1 - F(\tau_{j,K_j-1} - \mu_{ij})
 \end{aligned}$$

where $\mu_{ij} = x_i\beta_j$ and

- y_{ij} is the i -th country-year's score to item j (see Table 1)
- x_i is the latent level of democracy in country-year i
- the slope parameter β_j is the *item discrimination parameters*, tapping the extent to variation in the scores on the latent concepts generates different response probabilities; the 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.
- τ_j is a vector of unobserved thresholds for item j , of length $K_j - 1$.⁴
- $F(\cdot)$ is a function mapping from the real line to the unit probability interval. We use the familiar logistic CDF transformation, $F(z) = 1/(1 + \exp(-z))$.

This item-response model is very similar to a Bayesian factor analysis model; the similarities between the two models are elaborated in [Takane and de Leeuw \(1987\)](#) and [Reckase \(1997\)](#). The model presented here is similar to the logit/normit model of [Bartholomew and Knott \(1999, ch. 5\)](#). A chief advantage of Bayesian factor analysis or IRT models is that the latent variables, x , are modeled directly and (subject to identifying restrictions) are estimable and subject to inference just like any other model parameter. Contrast factor analysis, where the latent variables are referred to as *factor scores*, and are almost always treated as by-products of fitting a factor structure to the correlation matrix of the observed indicators: conventional factor analysis does not usually impose restrictions necessary to

⁴Note that our model for μ_{ij} does not have an intercept term *per se* and so we fit $K_j - 1$ thresholds; an alternative parameterization is to estimate $K_j - 2$ thresholds and an intercept, setting the first threshold to zero. These distinction between these two parameterizations is trivial; our parameterization is chosen for computational convenience.

uniquely recover factor scores, and hence there are multiple proposals for obtaining factor scores from factor analysis. This is an important distinction between the two measurement models: most applications of factor analysis are a models for the covariance structure of the indicators, rather than latent variable models *per se*, while the item-response approach models the observed indicators as functions of the latent variables.

In addition, classical factor analysis is not well suited for ordinal indicators. The standard factor analytic model is driven by assuming that conditional on the latent scores, the observed indicators have a multivariate normal distribution. This assumption does not hold in the present case where we have a series of ordered, categorical indicators. Asymptotically-valid corrections are available when working with discrete indicators, based on factor analysis of the polychoric correlation matrix of the indicators⁵ instead of the usual Pearson product cross-moment correlation matrix of the indicators. This highlights another strength of the item-response model: we can model the indicators on their own terms (binary, ordinal, multinomial, count, continuous, etc) without having to “correct” for the fact that our data aren’t of the type anticipated by the classical factor analytic model.

4.1 Identification and Estimation

The unknown parameters in our IRT model are $\boldsymbol{\theta} = \left\{ \underset{(n \times 1)}{\mathbf{x}}, \underset{(m \times 1)}{\boldsymbol{\beta}}, \mathbf{T} \right\}$ where $\mathbf{T} = (\boldsymbol{\tau}'_1, \dots, \boldsymbol{\tau}'_m)$. Polity IV is a large data set, with $n = 13,941$ country-years. Since we estimate a latent level of democracy for each country-year and we assume democracy to be unidimensional, we have (1) 13,941 latent variables to estimate, the x_i ; (2) $m = 5$ item discrimination parameters to estimate, the β_j , and (3) a number of threshold parameters for each of the five ordinal indicators, the $\boldsymbol{\tau}_j$. Given the data $\mathbf{Y} = \{y_{ij}\}$, we have the following likelihood for the data

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m \Pr(y_{ij}; x_i, \beta_j, \boldsymbol{\tau}_j)$$

where $\Pr(y_{ij}; x_i, \beta_j, \boldsymbol{\tau}_j)$ is defined above. Given the large number of parameters in $\boldsymbol{\theta}$, we work in a Bayesian setting, using Markov chain Monte Carlo methods to explore the joint posterior density of the model parameters. Details on the MCMC algorithm and our choice of priors appears below.

Note also that the model parameters are not identified without further assumptions or restrictions. Since the likelihood 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

⁵Olsson (1979) provides two methods of estimating polychoric correlations.

the β_j , yet provide the same likelihood. In particular, $\mu_{ij} = x_i\beta_j = x_i r\beta_j r^{-1}$ for any $r \neq 0$. More rigorously, we say that the model parameters are not *locally identified*, following the definition in [Rothenberg \(1971\)](#). For the special case of $r = -1$ we obtain a 180 degree rotation of the latent democracy scale, and we lack *global identification*. Further, the latent levels of democracy x_i can be all shifted by some constant c , yielding $\mu_{ij} = x_i\beta_j + c\beta_j$, with offsetting shifts in the threshold parameters τ_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. For our one dimensional model, global identification is trivial; given that we work in a Bayesian setting, the problem amounts to there being two “mirror-image” posterior modes and we simply choose the one that has the latent trait running from less democracy to more democracy as x_i increases.

As noted above, we work in a Bayesian framework so as to simplify estimation and inference for this model. In the Bayesian setting, interest centers on the joint posterior density of the model parameters, $\pi(\Theta|\mathbf{Y})$. Given the large number of parameters for this problem, the joint posterior density is extremely high dimensional. We use Markov-chain Monte Carlo methods to obtain a “random tour” of the parameter space supporting this joint posterior density; the virtue of MCMC methods is that (subject to regularity conditions) they produce random tours that visit locations in the parameter space 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 thus straightforward: we compute point estimates of latent levels of democracy by simply averaging the output of many iterations of the MCMC algorithm with respect to the x_i parameters (these averages are simulation-consistent estimates of the posterior mean). Assessments of the magnitude of the measurement error are obtained by computing the dispersion of the posterior density of each \mathbf{x} parameter; again, we compute this by noting, say, the standard deviation of the MCMC output with respect to the \mathbf{x} parameters.

Finally, the MCMC/Bayesian setting also has numerous computational advantages. 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 tremendously 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 $g_x(x_i|\boldsymbol{\beta}^{(t-1)}, \mathbf{T}^{(t-1)}, \mathbf{Y})$, $i = 1, \dots, n$
2. sample $\beta_j^{(t)}$ from $g_\beta(\beta_j|\mathbf{x}^{(t)}, \boldsymbol{\tau}_j^{(t-1)}, \mathbf{Y})$, $j = 1, \dots, m$

3. sample $\tau_j^{(t)}$ from $g_\tau(\tau_j|\mathbf{x}^{(t)}, \beta_j^{(t)}, \mathbf{Y}), j = 1, \dots, m$.

MCMC algorithms for ordinal response models are described in greater detail in [Johnson and Albert \(1999, 133-136\)](#). We implement this MCMC scheme using WinBUGS ([Spiegelhalter, Thomas and Best, 2000](#)), a general purpose computer program for estimation and inference via Bayesian simulation methods.

We employ normal priors for the discrimination parameters β_j with mean zero and variance 3^2 . After experimentation, we found that this prior variance is a reasonable choice for expressing ignorance yet helping keep the MCMC algorithm in numerically stable regions; since some of the Polity items discriminate extremely well with respect to the latent trait, more diffuse priors resulted in specific $\beta_j \rightarrow \infty$ and computational underflows when attempting to evaluate the logistic CDF extremely far into its tails. Our priors for the threshold parameters \mathbf{T} are also chosen so as to reflect prior ignorance tempered by the need for computational stability. We constrain the bottom threshold τ_{j1} to be no lower than -6 and the top threshold τ_{jK_j} to be no greater than 6, but otherwise specify nothing more than a simple ordering constraint: i.e.,

$$\begin{aligned} \tau_{j1} &\sim U(-6, \tau_{j2}) \\ \tau_{j2} &\sim U(\tau_{j1}, \tau_{j3}) \\ &\vdots \\ \tau_{jK_j} &\sim U(\tau_{jK_j-1}, 6) \end{aligned}$$

where $U(a, b)$ is the uniform distribution over the closed interval (a, b) . 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 let WinBUGS explore the posterior of the model parameters given the priors and likelihood described above, and impose the re-centering and re-scaling of the x_i on the output, iteration-by-iteration, effectively “post-processing” the MCMC output. Of course, transforming the x_i this way implies that the β_j and \mathbf{T} be appropriately transformed, and our estimates reported below are based on this transformation; thus the reported thresholds for some items lie outside the $(-6, 6)$ interval imposed by the prior. Examples of this “post-processing” approach for dealing with unidentified parameters via MCMC appear in [Hoff, Raftery and Handcock \(2002\)](#) and [de Jong, Wiering and Drugan \(2003\)](#).

We initialize the MCMC algorithm with start values of 1 for country years with Polity scores higher than 6, -1 for country years with Polity scores lower than -6, and zero otherwise. We initialize the β_j at 3 and the thresholds to be evenly spaced between -3 and 3. We let the algorithm run for 3,000 iterations as burn-in, moving away from the (arbitrary) start

values such that subsequent iterations represent samples from the joint posterior density. Estimates and inferences are based on 5,000 iterations, thinned by ten, in order to produce 500 approximately independent draws from the posterior density.

4.2 Latent Class Model

An ongoing dispute in the literature is whether democracy should be conceived as a continuum, or as a binary state. [Alvarez et al. \(1996\)](#) argues against the graded measures commonly employed, instead proposing a simple dichotomous classification of democracies and non-democracies. The various debates over continuum versus discrete measures are summarized in [Collier and Adcock \(1999\)](#). A researcher's commitment to a discrete conceptualization does not preclude using the approach described in this paper; the latent position of a country along a continuum defined by level of democracy can easily be replaced by membership in discrete classes.

In Latent Class Analysis, there are two sets of parameters to estimate: the probability of class membership (i.e., the proportion of democracies) and the conditional probability of a particular response to an indicator given class membership. The conditional probabilities are modeled as

$$\begin{aligned}
 \Pr(y_{ij} = 1 | D_i = 1) &= F(\tau_{j1}) \\
 \Pr(y_{ij} = 1 | D_i = 0) &= F(\tau_{j1} + \delta_j) \\
 \Pr(y_{ij} = 2 | D_i = 1) &= F(\tau_{j2}) - F(\tau_{j1}) \\
 \Pr(y_{ij} = 2 | D_i = 0) &= F(\tau_{j2} + \delta_j) - F(\tau_{j1} + \delta_j) \\
 &\quad \vdots \quad \quad \quad \vdots \\
 \Pr(y_{ij} = K_j | D_i = 1) &= 1 - F(\tau_{j,K_j-1}) \\
 \Pr(y_{ij} = K_j | D_i = 0) &= 1 - F(\tau_{j,K_j-1} + \delta_j)
 \end{aligned}$$

where τ_j is a vector of unobserved thresholds for item j , of length $K_j - 1$, and δ_j is the offset that distinguishes democracies from non-democracies.

The likelihood for each country-year's response pattern is a mixture of the multinomial

likelihood

$$f(\mathbf{y}_i) = \eta_i \prod_{j=1}^M \prod_{k=1}^{K_j} \Pr(y_{ij} = k | \text{Dem})^{Z_{ijk}} + (1 - \eta_i) \prod_{j=1}^M \prod_{k=1}^{K_j} \Pr(y_{ij} = k | \text{Not Dem})^{Z_{ijk}}$$

where $Z_{ijk} = 1$ if $y_{ij} = k$ and 0 otherwise. Let D_i be an indicator of latent class membership, i.e., $D_i = 1$ if i is a democracy and 0 otherwise. Then

$$\Pr(D_i = 1 | \mathbf{y}_i) = \frac{\theta \prod_{j=1}^M \prod_{k=1}^{K_j} \Pr(y_{ij} = k | \text{Dem})^{Z_{ijk}}}{f(\mathbf{y}_i)} \quad (1)$$

The estimated proportion of democracies is

$$\theta = \sum_{i=1}^N \Pr(D_i = 1 | \mathbf{y}_i) \quad (2)$$

The ordering constraints are imposed through inequality constraints on the cumulative probabilities as specified by Croon (1990). These inequality constraints are implemented by constraining δ_j to be negative, which implies $\Pr(y_{ij} = 1 | \text{Dem}) < \Pr(y_{ij} = 1 | \text{Not Dem})$ and $\Pr(y_{ij} = K_j | \text{Dem}) > \Pr(y_{ij} = K_j | \text{Not Dem})$. From each iteration from the Gibbs sample, for each country-year i , we obtain a binary indicator $T_i^{(t)}$, which is one if i is classified as a democracy at iteration t , and zero otherwise. Thus the mean of each T_i over many Gibbs samples is a simulation-consistent estimate of the probability in equation (1), and can be interpreted as a classification probability.

5 Analysis

We begin by briefly considering the question of dimensionality. After our recodings (see section 3, above), the five Polity IV indicators have the summary statistics and correlation matrix given in Table 2. We report Pearson correlations since they give essentially the same results as working with polychoric correlations for these data. The first eigenvalue is 3.61, while the second is less than 1.0, and via the “eigenvalues greater than one” rule of thumb, we would conclude that a uni-dimensional model is appropriate.

Table 3 presents the estimated discrimination parameters and the thresholds for each item. All of the items discriminate well with respect to the latent trait, with PARREG (Regulation

Marginal Distributions					
	XRCOMP	XROPEN	PARREG	PARCOMP	XCONST
1	.11	.11	.41	.28	.33
2	.54	.18	.28	.15	.05
3	.06	.08	.10	.30	.25
4	.29	.01	.18	.06	.02
5		.61		.17	.06
6					.02
7					.27
NA			.04	.04	
Mean	2.5	3.8	2.0	2.7	3.6
Std Dev	1.0	1.6	1.1	1.4	2.4

Pearson Product Moment Correlation Matrix					
	XRCOMP	XROPEN	PARREG	PARCOMP	XCONST
XROPEN	.67				
PARREG	.71	.39			
PARCOMP	.68	.36	.95		
XCONST	.75	.48	.72	.72	
Eigenvalues of Correlation Matrix					
	3.61	.81	.32	.19	.05

Table 2: Summary Statistics, Correlation Matrix, and Eigenvalues, Five Indicators from Polity IV

	Discrimination Parameter		Thresholds	
Competitiveness of Executive Recruitment (XRCOMP)	2.36 [2.29, 2.42]	τ_{11}	-3.46	[-3.54, -3.38]
		τ_{12}	0.90	[0.85, 0.94]
		τ_{13}	1.46	[1.40, 1.51]
Openness of Executive Recruitment (XROPEN)	1.40 [1.35, 1.45]	τ_{21}	-2.65	[-2.72, -2.59]
		τ_{22}	-1.19	[-1.24, -1.15]
		τ_{23}	-0.69	[-0.74, -0.65]
		τ_{24}	-0.63	[-0.68, -0.59]
Regulation of Participation (PARREG)	8.98 [8.76, 9.20]	τ_{31}	-2.26	[-2.36, -2.17]
		τ_{32}	4.10	[3.97, 4.23]
		τ_{33}	7.50	[7.40, 7.59]
Competitiveness of Participation (PARCOMP)	8.28 [8.15, 8.42]	τ_{41}	-4.59	[-4.64, -4.52]
		τ_{42}	-1.64	[-1.73, -1.55]
		τ_{43}	5.31	[5.20, 5.42]
		τ_{44}	7.39	[7.32, 7.46]
Executive Constraints (XCONST)	2.60 [2.54, 2.67]	τ_{51}	-1.48	[-1.53, -1.43]
		τ_{52}	-1.10	[-1.15, -1.06]
		τ_{53}	0.82	[0.77, 0.87]
		τ_{54}	0.99	[0.93, 1.03]
		τ_{55}	1.60	[1.55, 1.65]
		τ_{56}	1.84	[1.79, 1.90]

Table 3: Discrimination Parameters and Thresholds. Posterior Means, with 95% Highest Posterior Density Intervals in brackets.

of Participation) and PARCOMP (Competitiveness of Participation) providing extremely high discrimination and the threshold parameters for the extreme categories on these indicators pushing far out into the tails of the logistic CDF. The “Executive” variables have smaller item discrimination parameters, with XROPEN (Openness of Executive Recruitment) having the smallest of the five discrimination parameters. None of the five indicators are unrelated to the latent trait, but there is substantial variation in item discrimination: some of the Polity indicators tap the latent trait better than others, and any scale measure ought to reflect this.

Figure 1 shows the probability of the lowest and highest responses for each of the five Polity IV indicators, as a function of the latent trait; these curves are generated with the posterior means for the discrimination and threshold parameters reported in Table 3. The graphs highlight the extremely strong discrimination of the PARCOMP and PARREG items (the probability of any given coding on these indicators changes very abruptly over the range of the latent trait) and contrast the less pronounced discrimination of the indicators tapping aspects of executive competition, recruitment and constraints.

Figure 2 displays the distribution of the Polity scores and the IRT estimates (posterior means) for the entire 1800--2000 period, as well as for 2000 only. The Polity scores, especially for 2000, are heavily concentrated on the highest possible score, 10. In 2000, 34 of the 153 countries have perfect scores --- more than 20% of the countries. The high proportion of perfect scores suggests that there are either too few items and/or the difficulty of the items is low enough that the measure does not effectively distinguish among the observations (the analogy in the educational testing literature is administering a test that is too easy, and a large number of students receive perfect scores). The estimates based on our ordinal IRT model also show a distinct left-skew. In both the 2000 analysis and in the full 1800-2000 analysis, we find a large cluster of cases with extremely high levels of the latent trait, and relatively few cases with scores on the latent trait around 1.0 or below -1.0. These cases in the (most densely populated) upper tail and (less populated) lower tail are usually cases scoring high or low (respectively) on the high-discrimination PARREG and PARCOMP indicators. Because our analysis gives more weight to the information in these indicators than does the Polity scale, we push high or low scoring countries on these indicators relatively further into the tails on the resulting distribution of democracy.

In Figure 3 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 listwise deleting missing data); (3) the Polity IV scores themselves. For clarity and simplicity, we restrict the comparison to 2000 only. Figure 3 shows the three pairwise scatterplots among the three candidate measures in a matrix of a scatterplots; above the diagonal are the Pearson correlations among the three

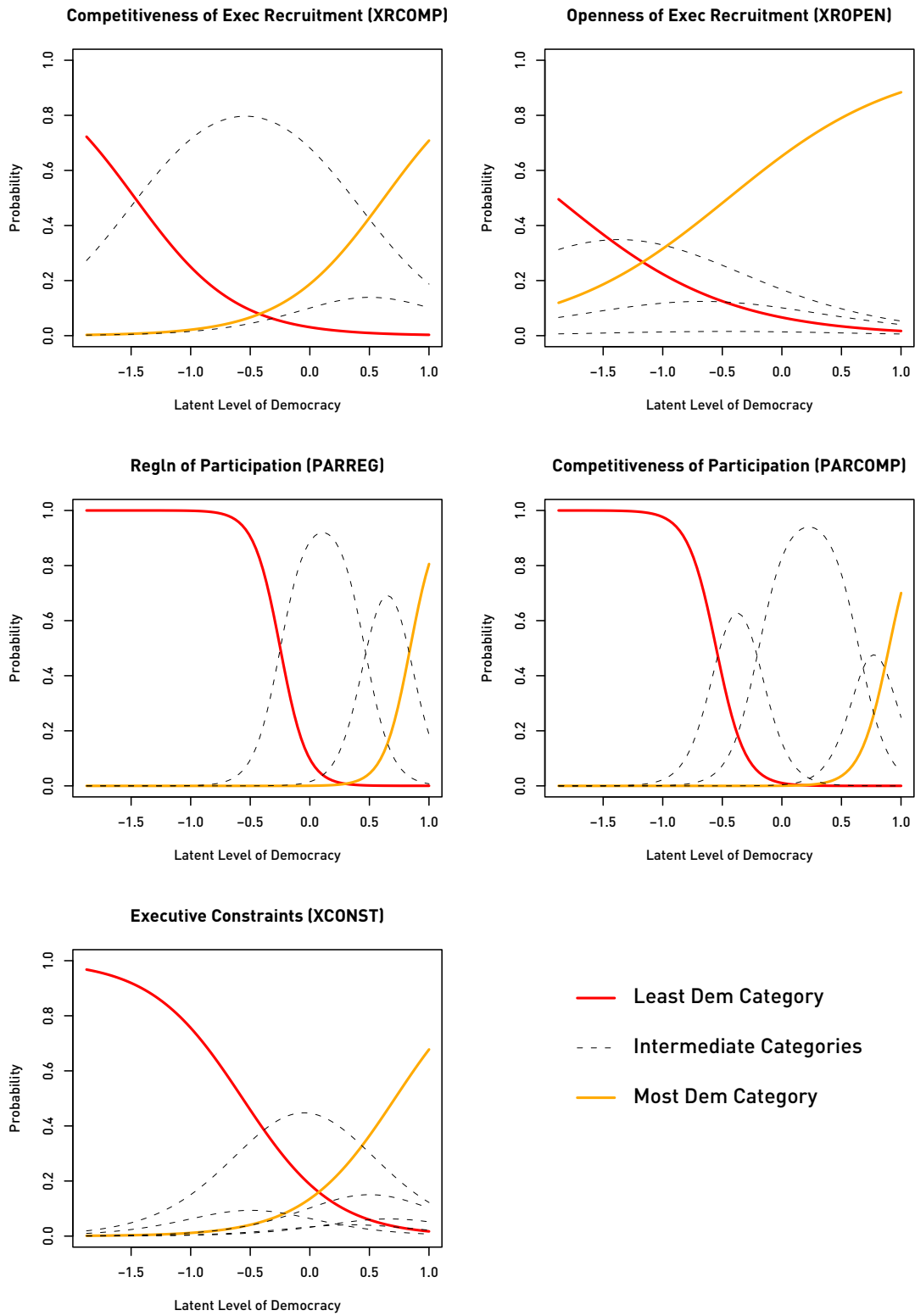


Figure 1: Item Characteristic Curves, Five Polity IV indicators

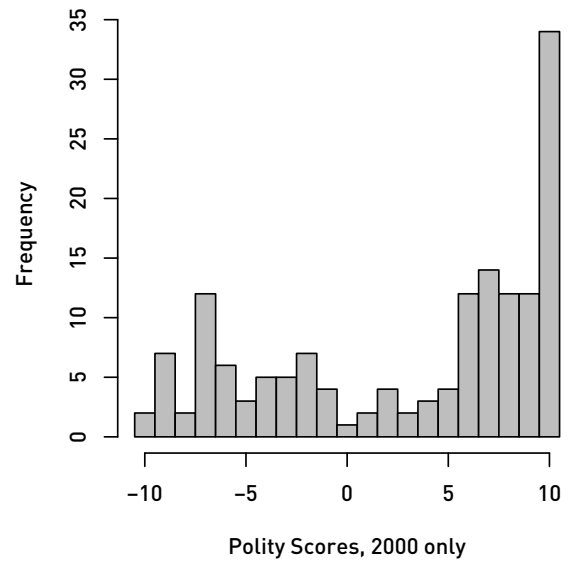
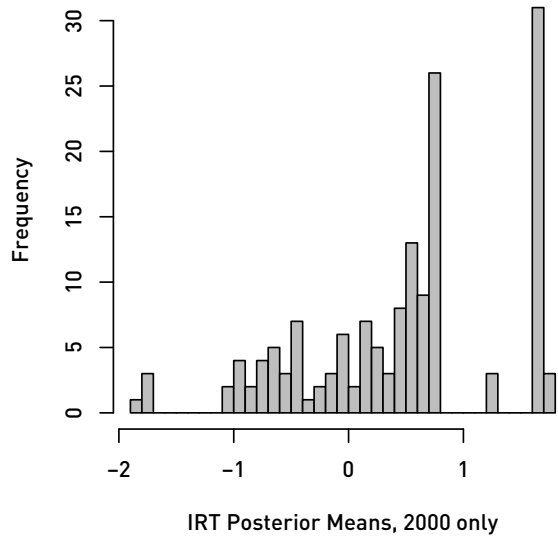
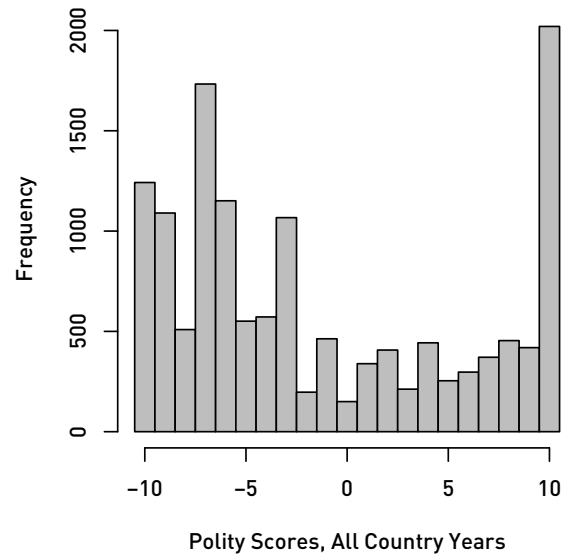
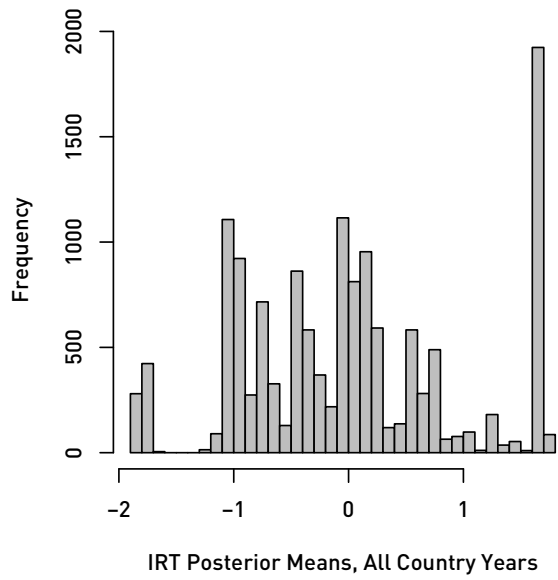


Figure 2: Distribution of Polity Scores and IRT Measures, 1800 - 2000 ($n = 13,491$) and 2000 only ($n = 153$).

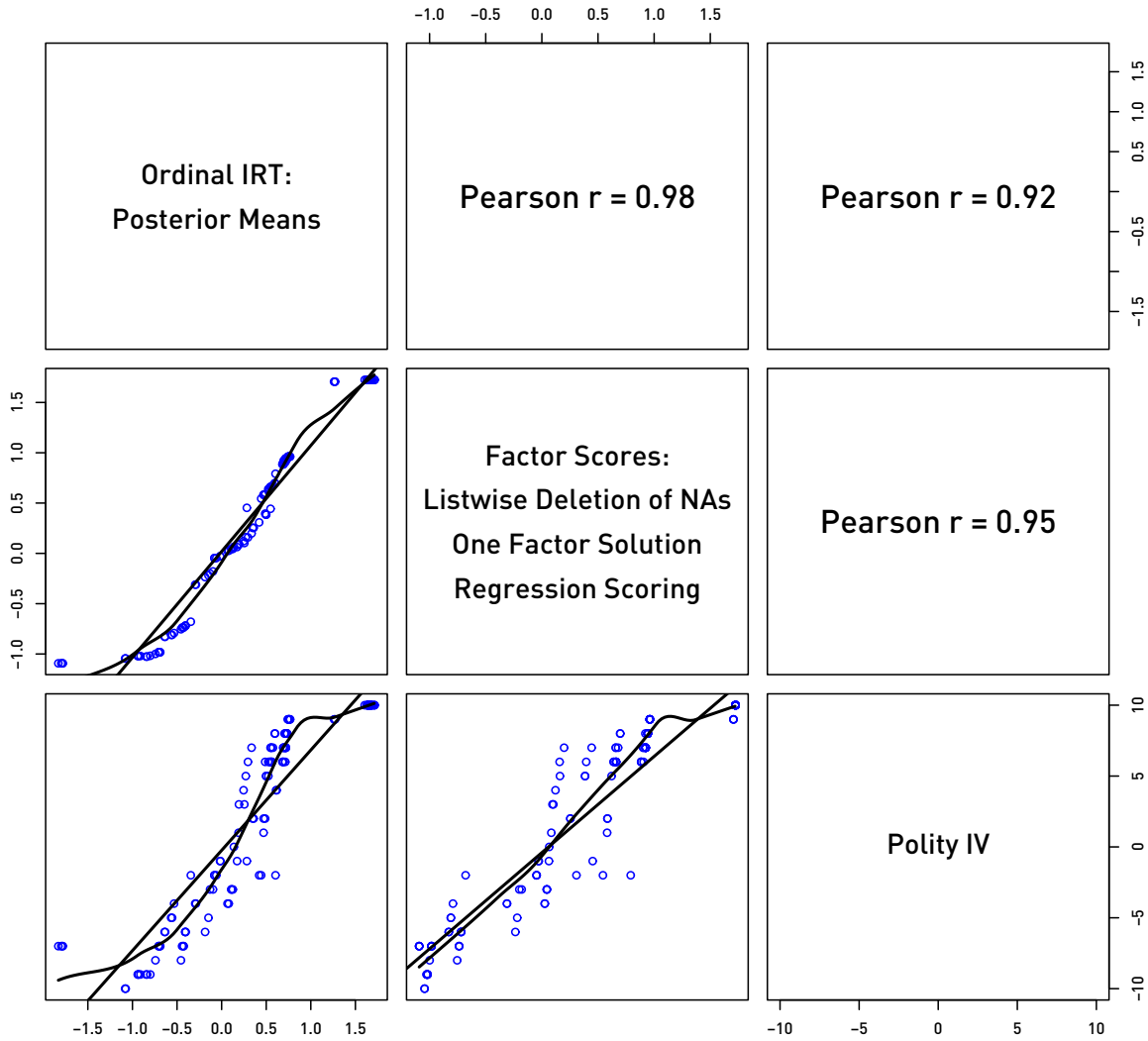


Figure 3: Comparison of Ordinal IRT Posterior Means, Factor Scores, and Polity, 2000 only. Lines for OLS local linear regression fits (span = 1/2, tri-cube kernel) are superimposed.

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 3) reflects the artificial “top-coding” in Polity: a score of 10 on Polity arises via the “maximum” response pattern 4-5-4-5-7). 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, reflecting some of the recoding we referred to in section 3, but largely stemming from the different weights our ordinal IRT model assigns to different indicators.

Ordinal IRT Model Latent Trait (Decile)	Median	Polity IV		
		2.5%	97.5%	Range
minimum - 10%	-7	-10	-3	7
10% - 20%	-9	-10	-9	1
20% - 30%	-7	-9	-5	4
30% - 40%	-6	-9	-2	7
40% - 50%	-3	-7	4	11
50% - 60%	-4	-6	4	10
60% - 70%	1	-3	6	9
70% - 80%	6	-4	9	13
80% - 90%	9	1	10	9
90% - maximum	10	10	10	0

Table 4: Comparison of Latent Trait (Posterior Means from Ordinal IRT Model) and Polity Score

In Table 4 we closely inspect the dispersion of Polity scores within each quantile 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 50%-60% range on our democracy scale could have a Polity score between -6 and 4, a range that covers half the 20 point Polity scale. Thus, if one were to treat our scores as “true scores”, then the Polity scores look somewhat unreliable.

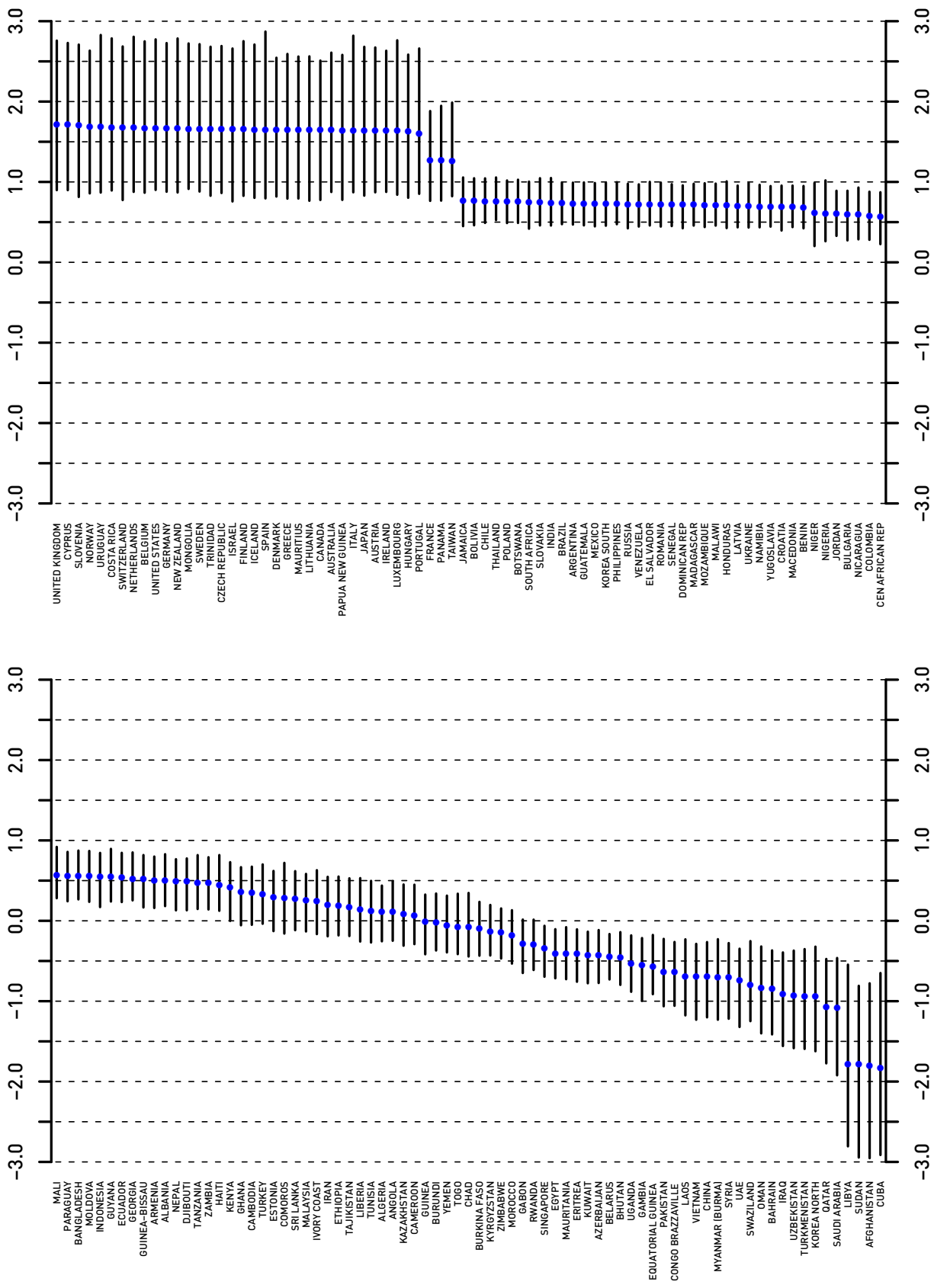


Figure 4: IRT Measures for 2000. Countries are ordered by their posterior means. Error bars indicate 95% highest posterior density regions.

5.1 Assessing Measurement Error in the Latent Trait

A key feature of our approach is the ability to easily compute and assess the measurement error in each country's latent level of democracy. Thus far, we have focused on our point estimates - the mean of the posterior distribution for each latent trait x_i . Now we turn attention to the posterior distributions themselves. In particular, the variance or dispersion of the posterior distribution of each x_i corresponds to our uncertainty in x_i after exploiting the information in the indicators and so reflects measurement error. And again, although we compute estimates covering the entire data period, for clarity and simplicity we concentrate primarily on the estimates for 2000 only.

Figure 4 displays the estimates for all 153 countries which received Polity IV codings in 2000. Unlike the Polity scores (or factor 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. The posterior density of each country-year's x_i is summarized with a point (the posterior mean) and a line covering a 95% confidence interval.

The striking feature of Figure 4 is that the measurement error increases in the extremes of the latent trait distribution. Countries that are estimated to be have either extremely high or extremely low levels of democracy also have substantially larger levels of measurement error. This is a familiar result when working with latent variable models. 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 we ask her: we know that the student is at the top of the class, but we until we see the student start to get items wrong, we can't put an upper bound on our estimate of the student's ability. Countries that get the maximum scores on the Polity indicators are like these students; we know that these countries are the most democratic in our data, but we do not get a precise estimate of the level of democracy in these countries. To do better we need more information, either in the way of more indicators or stronger theory: for instance, one might constrain the x_i of any country obtaining the maximum response profile (4-5-4-5-7) to some arbitrary value (e.g., 1.0, or 10.0), and perhaps likewise for the least democratic response profile (1-1-1-1-1). Of course, if we accept a position such as Dahl's - that no modern country is maximally democratic, that is theoretically possible that there are (as yet unobserved) more democratic scores to be had on the Polity indicators - then we also accept that the underlying scale is continuous and unbounded, and that while we can infer that the United States in 2000 is relatively democratic, we can't say exactly *how* democratic.

The consequences of measurement error are also apparent in Figure 5, which summarizes *differences* in estimates of each country's level of democracy from that of the United States,

again using the 2000 estimates only. For each country we can compute the difference between its x_i and that for the United States, $\delta_i = x_i - x_{US}$, and (of course), the uncertainty in that quantity. Figure 5 summarizes the posterior density of these δ_i quantities, with error bars covering a 95% confidence interval. A reasonably large set of countries have δ_i whose confidence interval overlaps zero, meaning that at a conventional 95% confidence level, they have democracy scores that are indistinguishable from that for the United States. For instance, our best guess is that the democracy scores for South Africa and the Ukraine in 2000 are almost one unit below the score for the United States in 2000, but because of measurement error, we fail to conclude that either country has less democracy than the United States at conventional levels of statistical significance.

In Figure 6 we graph the equivalent of a p -value for the one-sided hypothesis that $H_0 : \delta_i = x_i - x_{US} > 0$.⁶ Thirty-six or roughly one-quarter of the 153 countries available for analysis in 2000 have p -values greater than .05, implying that we can not distinguish their democracy score from that for the United States, at a conventional 95% level of statistical significance. Figure 6 also reveals that there does seem a pronounced break where we can distinguish cases from the United States, at least on a one-tailed test with a .05 significance level. There is a cluster of countries who have democracy levels essentially indistinguishable from that of the United States in 2000; these include the advanced, industrial democracies of the OECD and other countries such as Mongolia, Costa Rica, Trinidad and Papua New Guinea. But there are very few countries for whom we would have a difficult decision as to whether they were as democratic as the United States: only Taiwan, Panama and France fall into this category, with implied p -values of .27, .26 and .26 (i.e., the probability that these countries have a higher latent level of democracy than the United States). The next most dubious case in 2000 is South Africa; the probability that South Africa has a higher latent level of democracy than the United States in 2000 is just .024. And after that, there is very little doubt that the remaining countries are less democratic than the United States.

6 The Consequences of Measurement Error in Democracy

How consequential is measurement error in measures of democracy? That is, what inferential dangers are posed by using a measure of democracy in data analyses? As a general matter, it is well known that using “noisy” variables in data analysis generates an errors-in-variables problem that can potentially invalidate hypothesis tests. This problem is particularly pressing

⁶We 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.

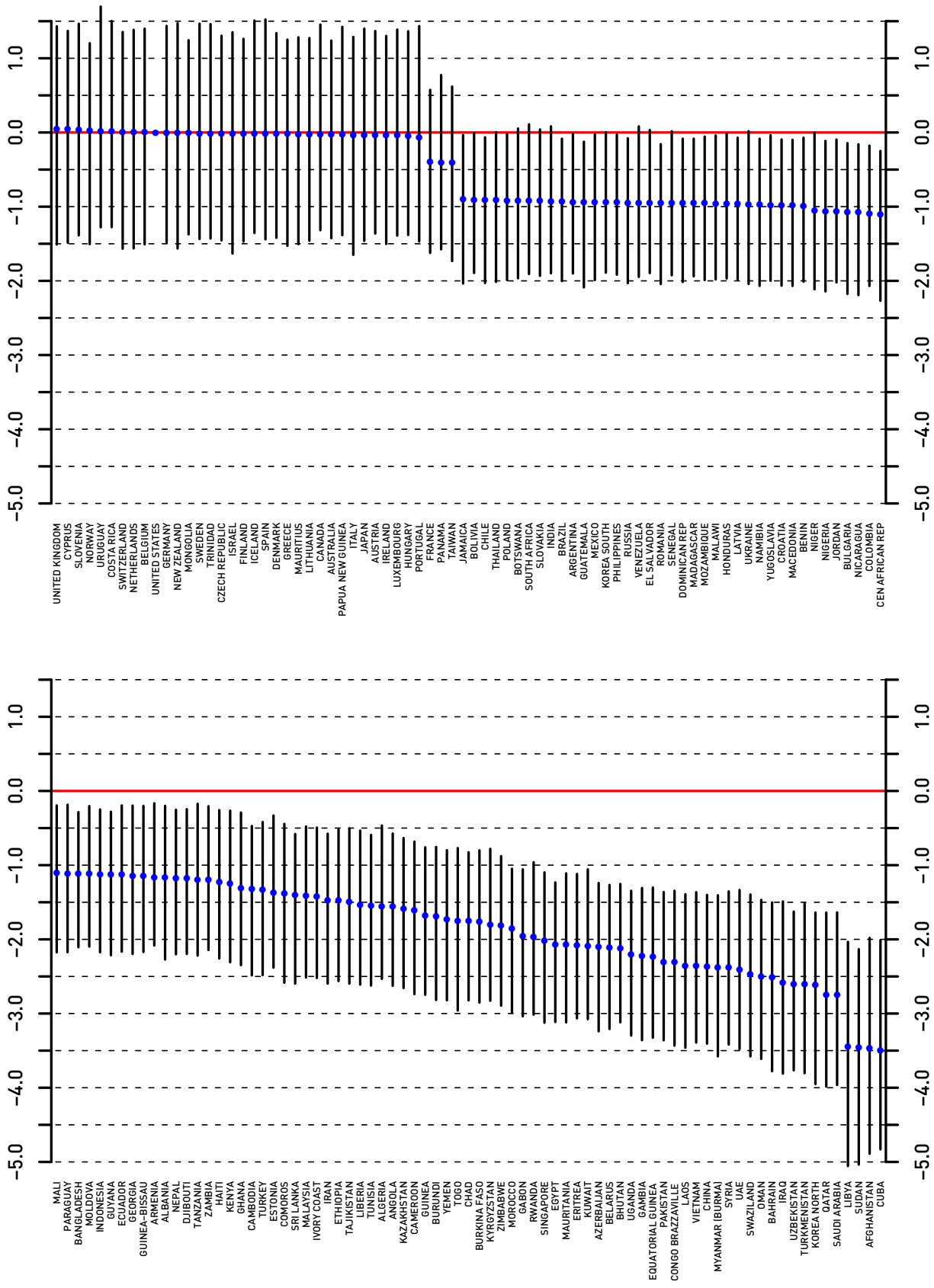


Figure 5: Difference from U.S. Posterior mean of difference between country measure and the score for the U.S., ordered by posterior means. Error bars are 95% highest posterior density regions.

when the variable measured with error appears on the right-hand side of a data analysis. For instance, a familiar result from intermediate econometrics is that if we regress y on X , where X is a noisy measure of ξ , then the estimated slope coefficient $\hat{\beta}$ is a biased and inconsistent estimate of the “true” relationship between y and ξ , and, in particular, is attenuated. Of course, this textbook result presumes that there exists a “true” ξ . In our approach we recover a country’s level of democracy as a posterior distribution, $\pi(x_i|y_i)$ summarizing what we believe about country i ’s level of democracy given the observable indicators y_i . The approach we adopt is to ensure that this uncertainty over the x_i propagates into subsequent analyses using x_i as a predictor variable.

6.1 Stylized Example

We use the following stylized example to make our point, before turning to two real examples from the literature. We consider an admittedly naïve analysis, where a country’s wealth (measured with the log of GDP per capita, measured in U.S. dollars, from the Penn World Table) is regressed on its level of democracy. We readily concede that this is not a rigorous statistical model of national wealth (see [Prezeworki et al. 2000](#), ch3 for an example of a more theoretically informed approach), nor would we wish to interpret the results as reflecting a causal relationship (since reasonable people differ as to whether democracy engenders national wealth, or vice-versa). Rather, our goal is to demonstrate the consequences of properly acknowledging the noise in the measure of democracy that we obtain from the Polity indicators.

We begin by regressing log GDP per capita on the democracy measure provided in the Polity data, for 117 countries with data available for both measures in the year 2000. Initial inspection of the data suggests that the relationship between these two variables is non-linear (see [Figure 7](#)), which is confirmed by a simple regression analysis; the r^2 of a quadratic regression fit to these data is .63, and the t -statistic for the quadratic term is 10 (see column one of [Table 5](#)) while the r^2 of the linear model is just .30. While we are reluctant to push hard on a substantive interpretation of the model, the non-linearity suggests that increases in democracy are not unambiguously associated with increasing national wealth; only at high levels of democracy is there a strong positive association between democracy and wealth, while at medium levels of democracy the association is weaker and is actually negative for levels of democracy below .22 (the 35th percentile of the Polity democracy scores for the year 2000). The non-linearity is in general suggestive of “transition costs”: we expect national wealth to decrease before increasing, as a country transitions from autocracy to democracy.

We replicate this analysis using the measure of democracy we obtain from the ordinal IRT

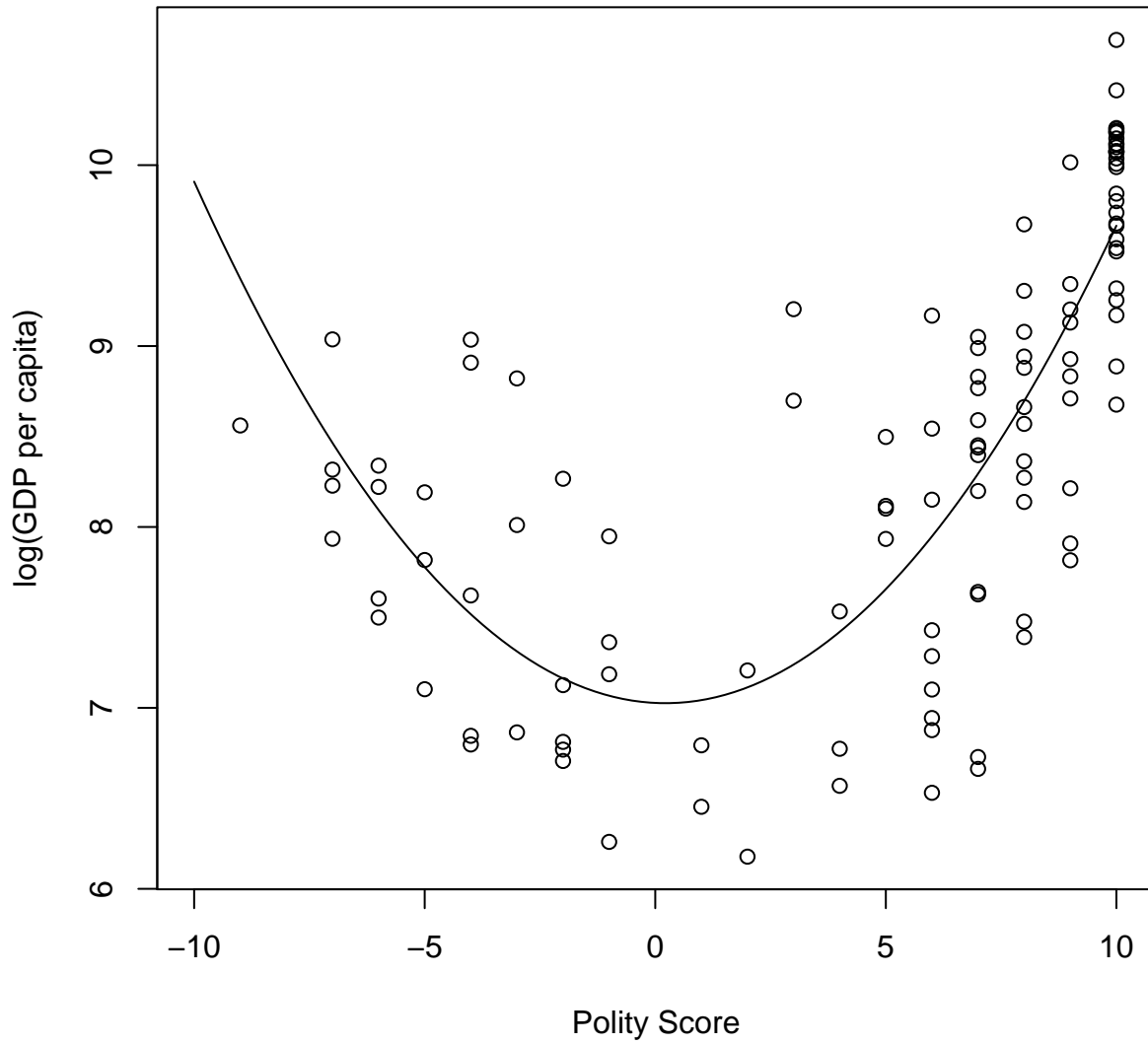


Figure 7: Relationship between Wealth (log GDP per capita) and Democracy (Polity score), 117 countries, measured in 2000.

model. First, we fit a naive model, ignoring the measurement error in our democracy measure by fitting the quadratic regression model using the posterior means of the democracy measure (our point estimates); estimates of this model appear in column two of Table 5. As for the Polity democracy score (column one), the data strongly support the quadratic specification and the same substantive implications discussed above.

	Polity Score	Measure of Democracy	
		Ordinal IRT Posterior Means Only	Ordinal IRT Sampling from Posterior
Intercept	7.0 [6.8, 7.3]	7.7 [7.5, 7.8]	7.9 [7.6, 8.1]
Linear Term	-.12 [-.045, .021]	.14 [-.25, .53]	.67 [.25, 1.1]
Quadratic Term	.028 [.022, .033]	.72 [.47, .97]	.16 [-.07, .42]
r^2	.63	.57	.40

Table 5: Regression Models, log of GDP per capita as a function of level of democracy. Intervals in brackets are 95% confidence intervals.

But the analyses in columns one and two of Table 5 rely on an especially strong assumption: that democracy is measured without error. We relax this assumption in computing the estimates reported in column three, with uncertainty in the democracy measures propagating into uncertainty assessments for the regression coefficients. Formally, in the Bayesian setting we adopt, we have the (conditionally Gaussian) regression model

$$z_i | x_i \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

with diffuse Normal priors for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and diffuse inverse-Gamma priors for σ^2 and where x_i is a random variable, with (posterior) density $\pi(x_i | y_i)$, and where z_i is log GDP per capita of country i . If x_i were known, then standard results apply, and via Bayes' Rule the posterior density of $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta} | \mathbf{z}, \mathbf{x})$ is multivariate normal with mean vector and variance-covariance obtained from simply running the regression on \mathbf{z} on \mathbf{x} (given diffuse priors). But since \mathbf{x} is a random variable (the uncertain product of a measurement procedure), we can't simply condition on it; instead we integrate over its predictive density, to give us the following

posterior density for β :

$$\pi(\beta|\mathbf{z}) = \int_{\mathcal{X}} \pi(\beta|\mathbf{z}, \mathbf{x})g(\mathbf{x})d\mathbf{x} \quad (3)$$

where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}$ and the predictive density $g(\mathbf{x})$ is actually the posterior density for \mathbf{x} from the measurement stage of the analysis, $\pi(\mathbf{x}|\mathbf{Y})$. Less formally, we are treating the democracy measures as akin to missing data, and use the ordinal IRT measurement model to supply multiple imputations when using level of democracy as an independent variable, with the extra uncertainty correctly propagating into inferences for the regression coefficients β .

We accomplish this via the following steps, performing the integration in equation (3) via Monte Carlo, also known as the “method of composition” (Tanner, 1996, 52). Letting t index samples from the posterior distribution for the democracy measures,

1. sample $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \dots, x_n^{(t)})'$ from the posterior distribution $\pi(\mathbf{x}|\mathbf{Y})$, where \mathbf{y} are the observed indicators of democracy in the Polity data set.
2. run the regression of log GDP per capita on $\mathbf{x}^{(t)}$, yielding $\hat{\beta}^{(t)}$ and $V(\hat{\beta}^{(t)})$.
3. sample $\beta^{(t)}$ from its posterior density, which (given an uninformative prior) is simply the multivariate normal distribution with mean vector $\hat{\beta}^{(t)}$ and variance-covariance matrix $V(\hat{\beta}^{(t)})$.

This process yields the results reported in column three of Table 5. The measurement error in the democracy measure is considerable, and results in a drop in model fit (r^2 falls from .60 to .43, column two to column three). But most importantly, the coefficient on the quadratic term is indistinguishable from zero, while the linear term becomes much larger (from .48 to .74, column two to column three). The substantive implications are quite striking: after we acknowledge the imprecision in the measure of democracy, more democracy results in more national wealth, at all levels of democracy. There is no longer any evidence for the “transition costs” interpretation of the quadratic model: that is, there is no longer any evidence to suggest that national wealth would decrease before increasing, as a country transitions from autocracy to democracy.

Why this occurs is reasonably straightforward. The cluster of countries in the upper-right corner of Figure 8 (those that score high on democracy and are relatively wealthy), are high leverage data points, dragging the fitted regression line towards it, providing support for the quadratic specification. But those countries are also the countries with high levels of measurement error, and their impact on the fitted regression is less pronounced when we properly acknowledge the measurement error, giving rise to the linear relationship shown in Figure 8.

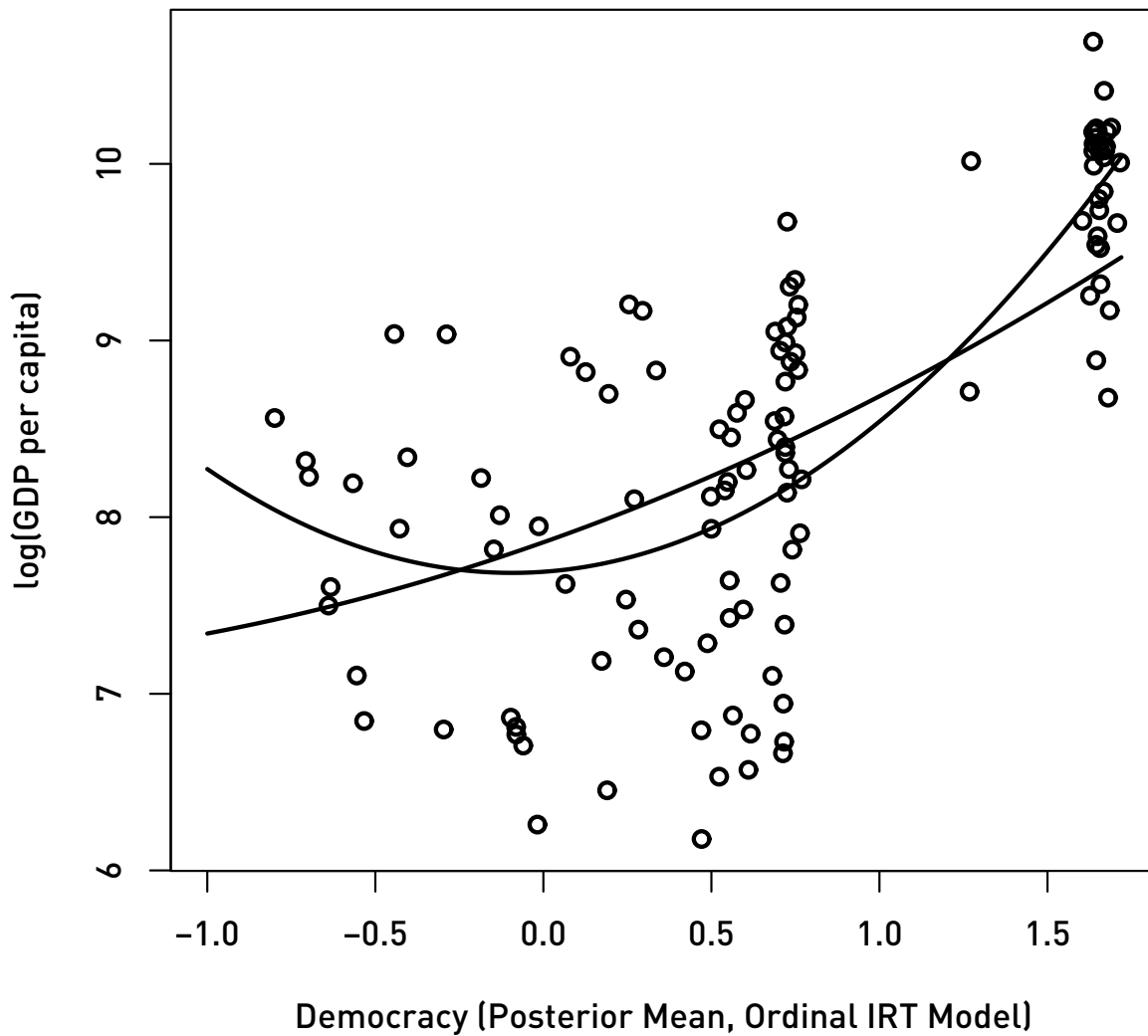


Figure 8: Mapping from Level of Democracy to National Wealth, With and Without Correction for Measurement Error in the Democracy Measure. The quadratic relationship (column two, Table 5) results when modeling with the point estimates of the democracy measure; the essentially linear relationship (column three, Table 5) results when the measurement error in democracy is properly accounted for.

7 Re-evaluating the Democratic Peace

A common use of the Polity scores is to test for the presence of a “Democratic Peace,” in which democracies are less likely to go war with each other. To demonstrate the consequences of aggregation and measurement error when replacing the Polity scores with our estimates, we use two examples: [Rousseau et al. \(1996\)](#) and [Oneal and Russett \(1997\)](#) (with the extensions of [Beck, Katz and Tucker \(1998\)](#) to handle duration dependence).

7.1 Monadic vs. Dyadic Democratic Peace

Rousseau et al. test the monadic and dyadic versions of the democratic peace argument. The monadic hypothesis identifies democracies as being less likely to resort to force in order to settle disputes, regardless of the nature of the regime in the opposing country. The dyadic hypothesis indicates that democracies are less likely to resort to force when the dispute involves other democracies.

The dataset consists of 303 dyadic conflicts adapted from the International Crisis Behavior (ICB) dataset. Each conflict appears twice in the dataset, once for each participant, for a total of 606 observations.⁷ For each observation in the dataset, the two participants are coded as the *Actor* and the *Opponent*, with the covariates corresponding to the Actor. The second 303 observations of dyadic conflicts correspond to the conflicts in the first half of the data, except that the labeling of Actor and Opponent are reversed. In our analysis, we use only 470 of the observations. The excluded observations include countries whose regimes are in transition, and for which Polity codes do not exist, and we did not include in the IRT analyses.⁸

The analysis involves ordered probits with two different dependent variables: Initiation of Force and Highest Level of Force. Both dependent variables contain three categories. Initiation of Force indicates the extent to which the state initiated conflict. For example, this variable is coded 2 if the state was the first participant to commit up to 1,000 troops to a combat zone (escalate the crisis to a minor level of force) and the first to commit more than 1,000 troops (escalate the crisis to a major level of force).⁹ The middle category indicates the country was the first to escalate to either a minor level or a major level, but not both, whereas the country is coded zero if they did not initiate force at all. Highest Level of Force measures the extent to which military forces were used. The country is coded 2 if at any point in the

⁷Rousseau et al. argue that in order to test the monadic/directed dyad hypotheses, the unit of analysis should be the nation-state, not the dyad.

⁸Rousseau et al. apparently coded the missing observations transitional states themselves using the Polity rules. Some of the observations were also colonies, which were included in Polity II but removed by Polity IV.

⁹This variable is probably not really ordered; it should be modeled as two dichotomous variables with one conditional on the other, or simultaneously determined.

crisis, it used more than 1,000 troops, coded 1 if less than 1,000 troops are used, and 0 if no troops were used.

The monadic and dyadic hypothesis are tested by the coefficients on Actor's Democracy, Opponent's Democracy, and an interaction between Actor's and Opponent's Democracy (Dyadic Democracy). Rousseau et al. use as their measure of democracy the scores from Polity II. They rescale the measures to range from 0 to 20. The monadic hypothesis expects the coefficient on Actor's Democracy to be negative, and the interaction effect to be zero. The dyadic argument hypothesizes that Actor's Democracy should have no effect, the coefficient on the interaction term should be negative, and the coefficient on Opponent's Democracy should be positive. Also included are three control variables: Balance of Forces, Shared Alliance Tie, and Satisfaction with the Status Quo. Balance of Forces is a continuous measure that ranges from zero to one, and the other two controls are dummy variables. The analysis for Highest Level of Force also includes the Opponent's Initiation of Force (the dependent variable in the first analysis).

The analyses are presented in Tables 6 and 7. Table 6 contains the analysis on Initiation of Force, and Table 7 contains the set of ordered probits on Highest Level of Force. The first column is a replication of Rousseau et al.'s analysis. Rousseau et al. concludes that the dyadic hypotheses are strongly supported. The interaction term and Opponent's Democracy are significant in both equations, at the 0.01 level, while Actor's Democracy misses the standard 0.05 cutoff with p-values of 0.064 and 0.094. The next two columns are the same set of analyses, except on the 470 observations that match the IRT estimates. The results are similar to the original estimation on all 606 observations. In the subsetted analysis, the effect of Actor's Democracy drops out completely, and the effect of balance of force decreases, but everything else remains the same.

The final two columns reestimate the Rousseau model by replacing the Polity measures with our continuous latent trait (IRT) and binary latent class estimates. The interaction term in their original paper is not the product of Actor's Democracy and Opponent's Democracy, but is the product of the Actor's Democracy measure, and a dichotomized measure of Opponent's Democracy. The dichotomized measure is coded one if the Opponent's Democracy is 17 or more, and zero otherwise. To replicate this, we dichotomize the Opponent's Democracy score by assigning the top 23 percentile as democracies and the bottom 77 percentile as non-democracies¹⁰ Using the latent trait scores, the effect of the interaction is completely gone (especially for the Initiate variable), Opponent's Democracy is no longer significant, and Actor's Democracy remains insignificant. Consequently, all of the effects are removed

¹⁰The 23 percent correspondent to the proportion of observations in Rousseau's Polity scores which are greater than or equal to 17.

when re-aggregating the Polity indicators and taking into account the measurement error, and neither the monadic nor the dyadic versions of the Democratic Peace can be confirmed.

The latent class analysis also discounts the dyadic hypothesis, where both Opponent's Democracy and the interaction are both statistically insignificant. Actor's Democracy though is significant ($p = 0.049$) in the Initiation of Force equation, and has a p -value of 0.066 in the Highest Level of Force specification. Thus, the LCA estimates indicate some evidence for the *monadic* hypothesis, directly opposite of Rousseau et al.'s conclusions.

The specification of the interaction term is a little arbitrary. Tables 8 and 9 reassess this analysis with a revised specification of the interaction term. The analyses from tables 6 and 7 are replicated except the interaction term is replaced by the product of Actor's Democracy and Opponent's Democracy. With respect to the original analysis using the Polity scores, all of the statistically significant effects relating to the democratic peace in the Initiation of Force analysis no longer hold. The p -values for the Highest Level of Force estimates drop slightly (except for Actor's Democracy, which was insignificant in the original analysis anyway), but the overall conclusions from Rousseau remain. Using the IRT estimates, the results reverse; there is weak support for a monadic version of the democratic peace, but not a dyadic version.

7.2 Liberal Democratic Peace

Oneal and Russett (1997) present another perspective on the search for a democratic peace. Their "liberal peace" supplements the traditional dyadic democratic peace with a dyadic economic peace: frequent trading parties are also less likely to go to war. Oneal and Russett's data ranges covers the period 1950-1985, and is constituted in the traditional dyad-year unit of analysis with 20,990 dyad-years. Beck, Katz and Tucker (1998) extend this analysis by accounting for temporal dependence using duration analysis. The dependent variable is the occurrence of a militarized dispute within the dyad. BKT implement the duration analysis using a standard logit model with either a series of dummies indicating the length of time since the the last dispute within the dyad, or a natural cubic spline of the number of years since the last dispute (or in the absence of disputes, years in the dataset). Of the independent variables, the democracy variable tests whether or not a democratic peace exists, and the coefficient on trade provides a test of the liberal economic peace. Democracy is measured as the minimum Polity score of the two dyad countries, and trade is defined as the lower bilateral trade-to-GDP ratio in the dyad.

Table 10 presents the original analysis of Oneal and Russett and Beck, Katz, and Tucker, as well as our IRT and LCA analyses. The first two columns use the Polity scores; the first column is the traditional logit specification of Oneal and Russett, and the second column

	Polity Scores		Estimates	
	Original Analysis	Subset Analysis	IRT Estimates	Latent Class
Intercept	-0.281 (0.175) [-1.611]	-0.247 (0.195) [-1.265]	-0.153 (0.141) [-1.087]	-0.064 (0.167) [-0.386]
Actor's Democracy	-0.0126 (0.0082) [-1.527]	-0.00049 (0.0093) [0.053]	-0.086 (0.076) [-1.14]	-0.258 (0.156) [-1.65]
Opponent's Democracy	0.027 (0.0099) [2.719]	0.026 (0.011) [2.342]	0.030 (0.064) [0.472]	-0.041 (0.144) [-0.282]
Interactive-Dyadic Democracy	-0.057 (0.019) [-3.02]	-0.069 (0.021) [-3.27]	-0.013 (0.167) [-0.080]	0.062 (0.064) [0.20]
Balance of Forces	0.996 (0.204) [4.872]	0.624 (0.234) [2.661]	0.670 (0.226) [2.96]	0.685 (0.229) [2.99]
Shared Alliances	-0.072 (0.145) [-0.493]	-0.080 (0.163) [-0.492]	-0.172 (0.161) [-1.070]	-0.167 (0.158) [-1.06]
Satisfied with Status Quo	-1.892 (0.181) [-10.466]	-1.874 (0.208) [-9.030]	-1.837 (0.210) [-8.749]	-1.818 (0.20) [-9.11]
Second Threshold	0.912 (0.066)	0.898 (0.074)	0.888 (0.073)	0.891 (0.069)
<i>N</i>	606	470	470	470

Table 6: Dependent Variable is Initiation of Force. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. All hypothesis tests are one-sided.

	Polity Scores		Estimates	
	Original Analysis	Subset Analysis	IRT Estimates	Latent Class
Intercept	0.080 (0.166) [0.48]	0.114 (0.185) [0.616]	0.130 (0.148) [0.874]	0.225 (0.159) [1.412]
Actor's Democracy	-0.0097 (0.0074) [-1.32]	-0.0041 (0.0083) [-0.49]	-0.057 (0.064) [-0.893]	-0.204 (0.135) [-1.506]
Opponent's Democracy	0.025 (0.0095) [2.69]	0.026 (0.011) [2.43]	0.0272 (0.058) [0.470]	0.0218 (0.138) [0.158]
Interactive-Dyadic Democracy	-0.070 (0.017) [-4.05]	-0.074 (0.019) [-3.88]	-0.129 (0.153) [-0.840]	-0.326 (0.321) [-1.02]
Balance of Forces	0.785 (0.093) [4.24]	0.577 (0.213) [2.71]	0.635 (0.200) [3.169]	0.649 (0.216) [3.01]
Shared Alliances	-0.240 (0.13) [-1.796]	-0.296 (0.152) [-1.95]	-0.349 (0.153) [-2.284]	-0.352 (0.148) [-2.38]
Satisfied with Status Quo	-1.96 (0.129) [-9.24]	-1.267 (0.144) [-8.78]	-1.223 (0.141) [-8.740]	-1.26 (0.15) [-8.62]
Opponent's Initiation of Force	0.87 (0.079) [10.94]	0.917 (0.090) [10.22]	0.893 (0.089) [10.03]	0.897 (0.095) [9.46]
Second Threshold	1.032 (0.064)	1.122 (0.076)	1.09 (0.073)	1.106 (0.078)
<i>N</i>	606	470	470	470

Table 7: Dependent Variable is Highest Level of Force. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. All hypothesis tests are one-sided.

	Original Analysis	Subset Analysis	IRT Estimates
Intercept	-0.241 (0.185) [-1.308]	-0.209 (0.206) [-1.015]	-0.163 (0.145) [-1.12]
Actor's Democracy	-0.0089 (0.013) [-0.674]	0.0082 (0.015) [0.566]	-0.096 (0.062) [-1.55]
Opponent's Democracy	0.016 (0.012) [1.38]	0.016 (0.013) [1.207]	0.023 (0.067) [0.35]
Interactive-Dyadic Democracy	-0.0012 (0.0013) [-0.924]	-0.0019 (0.0014) [-1.332]	-0.045 (0.067) [-0.67]
Balance of Forces	1.040 (0.203) [5.136]	0.651 (0.232) [2.805]	0.673 (0.235) [2.86]
Shared Alliances	-0.111 (0.144) [-0.768]	-0.127 (0.162) [-0.786]	-0.137 (0.160) [-0.856]
Satisfied with Status Quo	-1.876 (0.180) [-10.431]	-1.853 (0.206) [-9.006]	-1.824 (0.195) [-9.34]
Second Threshold	0.902 (0.065)	0.885 (0.073)	0.886 (0.073)
<i>N</i>	606	470	470

Table 8: Dependent Variable is Initiation of Force. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. All hypothesis tests are one-sided.

	Original Analysis	Subsetting Analysis	IRT Estimates
Intercept	-0.076 (0.18) [0.433]	0.094 (0.195) [0.483]	0.136 (0.15) [0.903]
Actor's Democracy	0.0022 (0.011) [0.20]	0.012 (0.013) [0.922]	-0.096 (0.063) [-1.527]
Opponent's Democracy	0.0197 (0.011) [1.77]	0.022 (0.012) [1.793]	0.0237 (0.064) [0.37]
Interactive-Dyadic Democracy	-0.00025 (0.0011) [-2.23]	-0.0031 (0.0013) [-2.447]	-0.040 (0.065) [-0.618]
Balance of Forces	0.841 (0.18) [4.57]	0.620 (0.212) [2.925]	0.635 (0.214) [2.96]
Shared Alliances	-0.265 (0.133) [-1.998]	-0.322 (0.151) [-2.138]	-0.347 (0.143) [-2.42]
Satisfied with Status Quo	-1.183 (0.129) [-9.191]	-1.266 (0.144) [-8.794]	-1.23 (0.14) [-8.79]
Opponent's Initiation of Force	0.841 (0.078) [10.82]	0.895 (0.088) [10.135]	0.891 (0.091) [9.81]
Second Threshold	1.019 (0.063)	0.885 (0.073)	1.10 (0.073)
<i>N</i>	606	470	470

Table 9: Dependent Variable is Highest Level of Force. The interaction for dyadic democracy is continuous. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. All hypothesis tests are one-sided.

	Polity Scores			IRT Estimates			LCA Estimates		
	Ordinary	Logit	Spline	Ordinary	Logit	Spline	Ordinary	Logit	Spline
	Logit			Logit			Logit		
Intercept	-3.29 (0.08) [-41.6]	-0.966 (0.09) [-10.45]		-3.09 (0.07) [-41.9]	-0.82 (0.09) [-9.40]		-2.73 (0.0118) [-23.2]	-0.53 (0.13) [-3.96]	
Democracy	-0.50 (0.07) [-6.73]	-0.54 (0.08) [-6.85]		-0.146 (0.049) [-2.97]	-0.28 (0.054) [-5.27]		-0.225 (0.082) [-2.76]	-0.0989 (0.089) [-1.109]	
Economic Growth	-2.23 (0.85) [-2.62]	-1.155 (0.92) [-1.26]		-1.76 (0.96) [-1.83]	-0.48 (1.03) [-0.47]		-2.35 (0.93) [-2.52]	-0.944 (1.05) [-0.899]	
Alliance	-0.82 (0.08) [-10.26]	-0.47 (0.09) [-5.25]		-0.88 (0.08) [-11.41]	-0.50 (0.09) [-5.41]		-0.85 (0.081) [-10.41]	-0.51 (0.087) [-5.83]	
Contiguous	1.31 (0.08) [16.50]	0.69 (0.09) [7.80]		1.37 (0.08) [16.61]	0.69 (0.085) [8.13]		1.47 (0.085) [17.38]	0.79 (0.0854) [9.27]	
Capability Ratio	-0.31 (0.04) [-7.43]	-0.30 (0.04) [-7.40]		-0.39 (0.04) [-7.10]	-0.30 (0.04) [-7.90]		-0.297 (0.044) [-6.78]	-0.306 (0.042) [-7.23]	
Trade	-66.13 (13.44) [-4.92]	-12.88 (10.51) [-1.23]		-104.5 (16.6) [-6.48]	-30.05 (12.94) [-2.32]		-128.8 (16.13) [-7.98]	-55.54 (13.96) [-3.98]	
N	20990	20990		20831	20831		20831	20831	

Table 10: BKT analysis. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. The coefficients from the cubic natural spline in Logit III have been omitted.

is Beck, Katz, and Tucker's update controlling for duration dependence. The standard logit analysis indicates that there exists both a politically oriented democratic peace, as well as a liberal economic peace. The BKT reconfirms the presence of Democratic Peace, as the effect of democracy are not attenuated after incorporating duration dependence. The coefficient on trade though is extremely attenuated, losing statistical significance as the z-statistic falls from -4.92 to -1.23. Thus, the hypothesis of a liberal economic peace is rejected.

Columns three and four replace the Polity scores with the minimum IRT estimates in each dyad. Level of democracy remains statistically significant at the 0.01, although the z-statistic for both analyses drop substantially. In particular, the z-statistic for the Oneal and Russett analysis drops more than half, from -6.73 to -2.97. The consequences of the measurement error is not limited to the effects of dyadic democracy on conflicts. Most importantly, the effect of trade is now statistically significant in both specifications, which would reaffirm Oneal and Russett's claim of a liberal democratic peace. The LCA results produce even stronger effects for trade, while the effect of the level of democracy no longer holds in the BKT model using the LCA classification.

In addition to the complications due to temporal dependence, another problem with the conflict data is the appearance of multiple "failures." Tables 11 and 12 contain BKT's various circumventions of the multiple failure problem, and our re-estimation. Table ?? removes all continuing dyad conflicts, since these are not really new instances of militarized dispute, but the same dispute spanning multiple years. Similar to the results in table 10, democracy as measured by the IRT estimates are statistically significant, but the z-statistics are attenuated. The only difference is the relationship between trade and militarized disputes is no longer significant when using the IRT estimates. The relationship between trade and disputes remains though when using the LCA estimates.

In table ??, the first three columns handle multiple failures by only considering the first dispute and discarding the remainder of the data. The last three columns differentiate multiple disputes by allowing the hazard to change stepwise according to a counter of the number of prior disputes. Again, both the democracy and trade hypotheses remain significant (albeit weakly in some of the equations); the exception is the effect of democracy washes out with the LCA estimates combined with a counter of prior disputes.

Consequently, once the Polity indicators have been re-aggregated, and the measurement error in the democracy variable is taken into account, the effect of democracy is attenuated, but largely remains. Furthermore, by instituting a proper measurement model for the level of democracy, the trade effect is once again significant, reconfirming Oneal and Russett's original claim. While these results largely hold whether or not the underlying theoretical concept of democracy is discrete or continuous, there are some differences based on the

	Polity Scores		IRT Estimates		LCA Estimates	
	Original Analysis	Duration Dependence	Original Analysis	Duration Dependence	Original Analysis	Duration Dependence
Intercept	-4.33 (0.11) [-37.9]	-3.57 (0.17) [-20.4]	-4.19 (0.010) [-41.2]	-3.45 (0.179) [-19.29]	-3.75 (0.17) [-22.2]	-3.03 (0.223) [-13.6]
Democracy	-0.40 (0.10) [-3.99]	-0.39 (0.10) [-3.87]	-0.120 (0.064) [-1.87]	-0.167 (0.066) [-2.52]	-0.306 (0.119) [-2.57]	-0.316 (0.118) [-2.68]
Economic Growth	-3.43 (1.25) [-2.74]	-4.01 (1.25) [-3.21]	-2.92 (1.30) [-2.19]	-3.35 (1.26) [-2.67]	-4.06 (1.34) [-3.036]	-4.46 (1.39) [-3.19]
Alliance	-0.48 (0.11) [-4.26]	-0.37 (0.11) [-3.20]	-0.509 (0.11) [-4.57]	-0.374 (0.012) [-3.04]	-0.481 (0.119) [-4.03]	-0.354 (0.138) [-3.109]
Contiguous	1.35 (0.12) [11.21]	0.99 (0.12) [8.02]	1.39 (0.125) [11.14]	1.00 (0.129) [7.74]	1.53 (0.132) [11.62]	1.149 (0.124) [9.27]
Capability Ratio	-0.20 (0.05) [-4.05]	-0.22 (0.05) [-4.56]	-0.19 (0.047) [-4.106]	-0.228 (0.049) [-4.669]	-0.184 (0.0439) [-4.19]	-0.218 (0.0517) [-4.22]
Trade	-21.08 (11.30) [-1.87]	-3.81 (9.68) [-0.39]	-47.15 (14.60) [-3.23]	-19.23 (12.66) [-1.52]	-64.13 (14.82) [-4.33]	-37.56 (13.78) [-2.73]
N	20448	20448	20298	20298	20298	20298

Table 11: BKT excluding continuing conflicts. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. The coefficients from the cubic natural spline have been omitted.

	First Disputes			Prior Disputes		
	Polity Scores	IRT Estimates	Latent Class	Polity Scores	IRT Estimates	LCA Estimates
Intercept	-3.21 (0.21) [-15.64]	-3.10 (0.20) [-15.35]	-2.65 (0.281) [-9.44]	-1.60 (0.10) [-15.52]	-1.49 (0.099) [-15.11]	-1.34 (0.148) [-9.02]
Democracy	-0.46 (0.13) [-3.47]	-0.176 (0.094) [-1.87]	-0.33 (0.17) [-1.95]	-0.41 (0.08) [-4.96]	-0.219 (0.059) [-3.72]	-0.028 (0.096) [-0.30]
Economic Growth	-2.29 (1.78) [-1.29]	-1.16 (1.80) [-0.647]	-2.01 (2.04) [-0.982]	-2.09 (0.97) [-2.16]	-1.37 (1.01) [-1.36]	-1.84 (1.03) [-1.78]
Alliance	-0.42 (0.16) [-2.59]	-0.452 (0.174) [-2.60]	-0.467 (0.158) [-2.94]	-0.25 (0.09) [-2.65]	-0.277 (0.094) [-2.94]	-0.29 (0.093) [-3.16]
Contiguous	1.11 (0.17) [6.44]	1.16 (0.176) [6.55]	1.33 (0.179) [7.39]	0.69 (0.09) [7.32]	0.69 (0.10) [6.76]	0.76 (0.095) [8.04]
Capability Ratio	-0.19 (0.06) [-3.42]	-0.180 (0.053) [-1.47]	-0.175 (0.053) [-3.27]	-0.20 (0.04) [-5.42]	-0.196 (0.037) [-5.30]	-0.199 (0.0364) [-5.48]
Trade	-3.55 (11.73) [-0.30]	-29.55 (20.12) [-2.77]	-54.11 (21.65) [-2.50]	-9.39 (10.19) [-0.92]	-24.56 (13.15) [-1.867]	-40.30 (13.23) [-3.05]
Prior Disputes				0.17 (0.01) [19.1]	0.17 (0.01) [18.52]	0.18 (0.0087) [20.41]
N	16991	16878	16878	20990	20831	20831

Table 12: Alternative methods of handling multiple failures: first disputes only, and a counter for prior disputes. Table entries are coefficient estimates, with standard errors in parentheses, and z-statistics in brackets. The coefficients from the cubic natural spline have been omitted.

conceptualization of democracy. In particular, using the LCA estimates, we also confirm the liberal economic peace hypotheses, but reject the political democratic peace arguments in several of the specifications. The IRT estimates, while always supporting the democratic peace hypothesis, provides only weak support for the liberal economic arguments in tables 11 and 12.

8 Conclusion

The Polity data have near-canonical status in political science; they provide an authoritative source of data on democracy that 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 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 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 are 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 a simple cross-national regression of log GDP per capita on democracy, we find that an apparently quadratic relationship is actually linear, 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, each of which is an ordinal item taking on four to seven levels. 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 ten to twenty self-placement items (each with seven 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). Adding even a few more indicators could improve the reliability of democracy measures considerably. 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). This design 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. This would be one way of expanding the amount of data available for measuring democracy.

Our second recommendation follows from the analysis in the second part of the paper. It is true that a better measure of democracy is a scientific advance in and of itself. But it is even more important to consider the consequences of working with a necessarily imperfect measure of democracy. The methodology we present in this paper provides a simple recipe for avoiding the over-optimism that can result when working with noisy measures. Failing to properly acknowledge the measurement in latent constructs risks inferential errors; scholars finding significant impacts of democracy on various dependent variables may well be wrong or (at least) over-stating 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 the ones we deploy here, ensuring that inferences about the effect of democracy on one's y 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 we create for the purposes of our data analyses: the tools we present here let us stop pretending otherwise.

References

- Alvarez, Mike, José Antonio Cheibub, Fernando Limongi and Adam Przeworski. 1996. "Classifying Political Regimes." *Studies in Comparative International Development* 31:3-36.
- Bartholomew, David J. and Martin Knott. 1999. *Latent Variable Models and Factor Analysis*. Arnold Publishers.
- Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42:1260--1288.

- Bollen, Kenneth A. 1980. "Issues in the Comparative Measurement of Political Democracy." *American Sociological Review* 45:370--390.
- Bollen, Kenneth A. and Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33:58--86.
- Bollen, Kenneth and Robert W. Jackman. 1989. "Democracy, Stability, and Dichotomies." *American Sociological Review* 54:612--621.
- Collier, David and Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science* 2:537--565.
- Croon, Marcel. 1990. "Latent Class Analysis with Ordered Latent Classes." *British Journal of Mathematical and Statistical Psychology* 43:171--192.
- Dahl, Robert. 1971. *Polyarchy*. Yale University Press.
- Dahl, Robert A. 1956. *A Preface to Democratic Theory*. Chicago: University of Chicago.
- de Jong, Edwin D., Marco A. Wiering and Mădălina M. Drugan. 2003. Post-Processing for MCMC. Technical Report No. UU-CS-2003-021. Institute of Information and Computing Sciences, Utrecht University.
- Eckstein, Harry and Ted Robert Gurr. 1975. *Patterns of Authority: A Structural Basis for Political Inquiry*. New York: Wiley-Interscience.
- Elkins, Zachary. 2000. "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations." *American Journal of Political Science* 44:287--294.
- Gasiorowski, Mark J. 1996. "An Overview of the Political Regime Change Dataset." *Comparative Political Studies* 29:469--483.
- Hewitt, Christopher. 1977. "The Effect of Political Democracy and Social Democracy on Equality in Industrial Societies: A Cross-National Comparison." *American Sociological Review* 42:450--64.
- Hoff, Peter, Adrian E. Raftery and Mark S. Handcock. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97:1090--1098.
- Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Late Twentieth Century*. Norman: University of Oklahoma Press.
- Johnson, Valen E. and James H. Albert. 1999. *Ordinal Data Modeling*. NY: Springer-Verlag.

- Linz, Juan J. 1975. "Totalitarian and Authoritarian Regimes." In *Handbook of Political Science*, ed. Fred Greenstein and Nelson Polsby. Vol. 3 Addison-Wesley pp. 175--353.
- Lipset, Seymour Martin. 1960. *Political Man: The Social Bases of Politics*. New York: Anchor.
- Marshall, Monty G. and Keith Jagers. 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 Jagers. 2002b. "Polity IV project: Political Regime Characteristics and Transitions, 1800--2000. The Polity IV Dataset." Retrieved from <http://www.cidcm.umd.edu/inscr/polity/>.
- Marshall, Monty G., Ted Robert Gurr, Christian Davenport and Keith Jagers. 2002. "Polity IV, 1800--1999: Comments on Munck and Verkuilen." *Comparative Political Studies* 35:40--45.
- Muller, Edward N. 1988. "Democracy, Economic Development and Income Inequality." *American Sociological Review* 53.
- Munck, Gerardo L. and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35:5--34.
- Olsson, Ulf. 1979. "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient." *Psychometrika* 44:443--460.
- Oneal, John R. and Bruce M. Russett. 1997. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950--1985." *International Studies Quarterly* 41:267--293.
- Powell, G. Bingham, Jr. 1982. *Contemporary Democracies*. Cambridge, Massachusetts: Harvard University Press.
- Prezeworki, Adam, Michael E. Alvarez, José Antonio Cheibub and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. New York: Cambridge University Press.
- Reckase, Mark D. 1997. "The Past and Future of Multidimensional Item Response Theory." *Applied Psychological Measurement* 21:25--36.
- Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39:577--591.
- Rousseau, David L., Christopher Gelpi, Dan Reiter and Paul K. Huth. 1996. "Assessing the Dyadic Nature of the Democratic Peace, 1918--88." *American Political Science Review*.

Rustow, Dankwart. 1967. *A World of Nations: Problems of Political Modernization*. Washington, DC: Brookings Institution.

Spiegelhalter, David J., Andrew Thomas and Nicky Best. 2000. *WinBUGS Version 1.3*. Cambridge, UK: MRC Biostatistics Unit.

Takane, Yoshio and Jan de Leeuw. 1987. "On the Relationship Between Item Response Theory and Factor Analysis of Discretized Variables." *Psychometrika* 52:393--408.

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