

A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence

Drew A. Linzer
Emory University
Department of Political Science
dlinzer@emory.edu

Jeffrey K. Staton
Emory University
Department of Political Science
jkstato@emory.edu

Abstract

We develop a measurement model for latent concepts commonly encountered in time series, cross-sectional analyses in comparative politics and international relations. Our approach addresses related patterns of missing data, measurement error and temporal dependence found in these data, which derive from the fact that manifestations of the underlying concept are not produced by the units themselves but rather by independent teams of scholars who have tried to measure the concept using more or less similar information. It thus enables scholars to more precisely link conceptual definitions to technical assumptions necessary for estimation. We use the model to generate a new time series, cross-sectional measure of judicial independence, which is available for 50 years and 200 states, and which solves several vexing problems that undermine the utility of existing indicators.

Prepared for presentation at the 2011 Annual Meeting of the American Political Science Association, September 1-4, Seattle, Washington.

Introduction

Most of the central concepts of comparative politics and international relations are latent. We cannot directly see a state’s “level of democracy,” “degree of corruption,” or “military resolve.” We may believe that we can meaningfully compare states along these dimensions, but it is not because we can directly observe the concepts. Instead, we appeal to features of the world that we can observe and believe to be related to the features we cannot. For this reason, the measurement of many concepts in IR and CP requires inference (Treier and Jackman, 2008, 203). Scholars often approach the challenge via a single proxy operationalization, as when democracy is measured by a regime score (e.g., Przeworski et al., 2000) or when the strength of a party is measured by its age (e.g., Stokes, 2001); however, doing so asks a lot of the proxy and is hard to justify when plausible alternative indicators exist. Yet using multiple indicators raises immediate questions about how to aggregate them sensibly and without substantial loss of information. Recalling that we confront an inferential challenge, the question is how to use the multitude of useful information available to us in a systematic and sensible way.

We develop a heteroscedastic graded response IRT model for time series, cross-sectional data, that is applicable to a variety of latent concepts in IR and CP.¹ Item response models have been most commonly applied in political science to measure legislator, judge or party ideology (Bailey, 2001; Martin and Quinn, 2002; Clinton, Jackman and Rivers, 2004; Voeten, 2007; Lauderdale, 2010); however, comparative scholars have more recently used the approach to scale regime-level concepts of democracy (Treier and Jackman, 2008; Pemstein, Meserve and Melton, 2010). We build on these efforts yet seek to more precisely tailor the approach to common features of quantitative comparative data. First and foremost, the model is designed to apply to TSCS data where we expect not only that the latent variable trends smoothly over time, but also that the observed data are not independently distributed within

¹The term “graded response” in the context of item response theory refers to models for outcome variables that are measured at the ordinal level (Johnson and Albert, 1999, ch. 7).

countries. Second, though it is standard to assume that the latent variable is continuous and unbounded, as we discuss below, core concepts in IR and CP have natural boundaries. Our model reflects this.

Third, we address implications of a core difference between the process by which, say, legislative votes and country-level indicators of democracy emerge. Standard scaling models were designed to uncover features of legislators, parties or students; and in such cases, the observable indicators of ideology or ability are produced directly by the units of interest. In the context of a model designed to synthesize multiple indicators for the purpose of scaling states, manifestations of the latent concept are not produced by the units but instead by teams of scholars looking to measure the underlying concept for each unit (Pemstein, Meserve and Melton, 2010, 431). A number of complications of measurement error and missing data emerge from this important difference in the data generating process. Most alarmingly, distinct research teams need not share the same conceptual definition of the latent concept. Even if they do, they may differ over what should constitute valid observable indicators. Relatedly, some research teams simply may be better at measuring some concepts, perhaps especially so in particular regions or political contexts. And of course, some states may be more difficult to measure than others. Finally, given different scopes of research, and perhaps even the capacities of the teams, missing data is far from random. In some cases, it is related to measurement error (e.g., Ríos-Figueroa and Staton, 2010). We develop a model that deals with these concerns in a theoretically coherent way.

We use our model to construct a time series, cross-sectional measure of judicial independence, which is available for 200 states and 50 years. In part, we focus on judicial independence because existing measures of the concept present the conceptual and theoretical challenges the model is designed to address. But we also apply the model to this context for substantive reasons. Judicial independence is central to multiple concerns with governance around the world, and although there is evidence suggesting that existing measures are valid, there are compelling reasons to seek a new measure that is available for a broader

cross-section of states and a longer period of time. In what follows, we further develop the measurement challenges we confront, discuss how our approach addresses these challenges, and specify the model. We then discuss why judicial independence reflects a particularly appropriate context in which to apply the model, describe our data and summarize the model's results. We conclude with a few remarks on the implications of this study for measurement in IR/CP generally and for judicial politics specifically.

Research Background

Quantitative comparative research often requires the use of a conceptual variable that can be described qualitatively but is not directly observable. There are generally three ways to proceed in such a case. First, we can appeal to a single proxy operationalization. As we discuss above, when multiple, plausibly good alternatives exist, the choice is not obvious. And unless we believe that measurement error is extremely small or that a particular indicator matches our theoretical needs particularly well, asking which indicator to use poses the wrong question. A second approach involves using a deterministic set of rules for systematically combining information from related variables into a single category or scale. One simple but often effective version of this technique is to average together a battery of distinct measures that are each believed to be manifestations of the underlying, latent concept (e.g., Ansolabehere, Rodden and Snyder, 2008). A variety of methods for eliciting and pooling expert ratings are also available (O'Hagan et al., 2006). The benefit to this approach is that the resulting scale will be directly interpretable as a measure of the latent concept—assuming that the aggregation procedure rests on a solid theoretical foundation.

In certain applications, however, the association between the latent variable and each of the manifest indicators is not sufficiently evident *a priori* to use such descriptive techniques. This may be because some manifest variables are better or worse indicators of the underlying scale; or because there is a large amount of measurement error or stochastic noise in the

observed indicators. It may also be the case that not enough data are available to reliably implement the aggregation procedure for the entire set of observations.

A third option is a statistical measurement model that specifies a mathematical function linking the latent concept to the manifest variables probabilistically, with parameters estimated from the data. These models are designed to allow the data to “reveal” their own underlying structure, including the relative importance of each observed indicator to the latent quantity. Scaling models such as the Rasch or item response model are examples of measurement models that infer a latent continuum from multiple categorical variables (van der Linden and Hambleton, 1997; Fox, 2010). Principal component and factor analytic models perform dimensional reduction, converting a series of continuous variables into a smaller number of scales that preserve common information (Jolliffe, 2002; Quinn, 2004; Arel-Bundock and Mebane, 2011). These methods offer certain advantages over more basic descriptive approaches, but they must also be justified and interpreted with care. It is not guaranteed, for example, that the latent dimension or categorization that is “found” by these techniques will match the theoretical concept originally being sought. To ensure that proper inferences are being drawn about the latent variable, researchers must choose manifest indicators well, and pay close attention to posterior validation, model fit, and the sensitivity of the results to the chosen model specification.

Aside from these general concerns, a measurement model designed for common concepts in IR and CP confronts a number of additional challenges. Comparative TSCS data—in which the same set of countries are measured repeatedly over time—share a set of features that further complicate the use of standard measurement models. Most latent variables in comparative research are meant to represent an underlying characteristic that changes in a relatively gradual and consistent manner from year to year. Manifest variables, in contrast, are expected to contain a greater amount of noise, and be much more prone to idiosyncratic yearly fluctuations. Applying measurement models that assume independent and identically distributed observations within countries (as most do) ignores the temporal element of these

data and does not necessarily provide for a smoothed latent trend. Simply averaging together multiple indicators will also not produce smooth year-to-year trends in the latent variable, unless the researcher has access to a large number of manifest variables that are consistently observed across most countries and years. The approach we advocate allows for smoothing over time.

Another conceptual consideration is that many latent variables in comparative research have natural lower and upper bounds. Income cannot be more equally distributed than on a perfectly equal basis, nor can it be less equally distributed than when a single person controls all of a state's resources. The ratio between the size of the winning coalition and the selectorate must lie on the unit interval (Bueno de Mesquita et al., 2003). A central bank or a judiciary can only be so dependent or independent of a sitting government. At some point, we should treat bankers and judges as either the government itself or, well, themselves. Our approach allows for reasonable bounds to be placed on the latent variable, thus allowing for a more precise match between conceptual definitions and their empirical manifestations.

It is also unlikely that countries are equally easy to measure. One natural possibility is that certain countries (perhaps during particular periods) are simply more poorly understood. This may be because the set of knowledgeable people about State A (e.g., Russia) is far larger than State B (e.g., Suriname). It may be because some countries collect and make available more (and potentially more detailed) information than others. Or it may be that a concept is more relevant to politics in one place than another. For example, it may be easier to measure "military resolve" in Israel or the United States than in Costa Rica, since the former countries are commonly involved in international conflicts, whereas Costa Rica does not have an army. Our approach addresses this concern by estimating the "measureability" of states.

From a more practical perspective, the availability of manifest indicators of various latent concepts in comparative data is highly uneven. There are typically few theoretically relevant manifest variables per latent variable to begin with—perhaps no more than ten or fifteen.

Those that do exist rarely span the entire range of countries or years under investigation. Missing data in most comparative indicators is extensive (Honaker and King, 2010). With such sparse data, a measurement model is preferred that will be robust to the presence of intervals when data are limited or nonexistent.

Model specification

We describe a latent variable measurement model for comparative time series cross-sectional data that estimates a smoothly trending value along a unidimensional, bounded interval scale. The latent variable x_{kt} varies across both countries $k = 1 \dots K$ and years $t = 1 \dots T$. Although x is unobserved, we assume there to be a series of R observed variables y^r , also measured at the country-year level, that can be taken as indicators, or ratings of the latent concept x . Individually, each y^r is an imperfect and incomplete measure of the latent concept; but together, the y^r are able to reveal variation in the level of x across countries and over time. Our model provides a statistical mechanism for combining the observed y_{kt}^r to produce reliable estimates of the underlying x_{kt} .

The indicators y^r that are used to assess x are chosen by the analyst on the basis of a prior theoretical expectation about the implications of larger or smaller values of x_{kt} in the world. Clearly, if we have access to multiple indicators that are specifically designed to capture a particular latent concept, then we are well-advised to use them. But we might also consider proxy indicators, especially those that our theories suggest are manifestations of the underlying concept. For example, states with higher levels of democracy may exhibit more frequent leadership change. States with relatively high levels of social capital may have larger participation (per capita) in amateur sports clubs. States with lower levels of judicial independence might exhibit politically motivated purges or a pattern of decision-making that is highly sensitive to government preferences. In none of these cases are the observable manifestations of the concept equivalent to the concept itself; however, we can have strong

theoretical or conceptual justifications for including them nonetheless. We assume only that x and y^r are positively associated (otherwise, treating y^r as a manifestation of x is likely incorrect); and that 1) certain indicators may be better or worse measures of the latent variable, and 2) certain countries may be more or less reliably measured by those indicators. This could be due to the inherent difficulty of learning about y^r in country k , or the possibility that the theory relating x to y^r does not apply equally well in all countries. We require no other micro-level assumptions about the *actual* process by which x_{kt} results in y_{kt}^r .

It is most common in quantitative comparative research for the y^r to be measured as ordinal-level scales. When manifest variables are recorded at the interval level, they are often bounded—from above, below, or both. As noted by Pemstein, Meserve and Melton (2010, 433), “although these scores take on many values and thus resemble interval scales, they do not necessarily provide interval-level information” about the latent variable. The primary concern is that the relationship between a continuous y^r and the latent x may not be linear. Following Pemstein, Meserve and Melton (2010), we therefore convert any continuous y^r into ordinal rankings. Variables containing more observations can be partitioned more finely into larger numbers of discrete categories, with minimal loss of information. This also preserves a consistent interpretation of model parameters across the set of manifest variables. The results of our analysis are robust to the categorization rule.

To link the latent x_{kt} to the manifest y_{kt}^r , we specify a heteroscedastic graded response IRT model. A series of item coefficients, β_r , capture the reliability, or *discrimination*, of indicator r as a measure of x . Treier and Jackman (2008, 205) describe this parameter as “the extent to which variation in the scores on the latent concepts generates different response probabilities” in the outcome variables. Larger estimates of β_r reveal a closer relationship between x and y^r . The inverse of this parameter has an equally intuitive interpretation as the “personal error variance” of rater r (Pemstein, Meserve and Melton, 2010, 431). A more “noisy” relationship between x and y^r is indicated by estimates of β_r that are closer to zero.

A second set of coefficients, γ_k , reflect the overall reliability of indicators y^r in country k . The γ_k act as a country-level multiplier for the β_r . If a country is harder to measure, or if the manifest variables are all more weakly associated with the latent variable, then smaller values of γ_k attenuate the effects of β_r . Countries in which the latent variable tracks more closely with the manifest variables will have larger γ_k . The idea is similar to the heteroscedastic IRT model developed by Lauderdale (2010), which applies when certain observational units respond more “unpredictably” to measurement by various manifest indicators. Since we assume that x and y^r are positively associated, we restrict both $\beta_r \geq 0$ and $\gamma_k \geq 0$.

Denote as M_r the total number of outcome categories for the r th manifest variable. Also let τ_{rm} represent the threshold values for item r in the graded response model, with $m = 1 \dots M_r$. The τ_{rm} divide adjacent ratings on the latent scale, subject to the constraint that $\tau_{rm} > \tau_{r(m-1)}$. Then the link function is written:

$$\Pr(y_{kt}^r = m) = \text{logit}^{-1} \beta_r \gamma_k (\tau_{rm} - x_{kt}) - \text{logit}^{-1} \beta_r \gamma_k (\tau_{r(m-1)} - x_{kt}). \quad (1)$$

We fix $\tau_{r0} = -\infty$ and $\tau_{rM_r} = \infty$, and estimate the remaining $M_r - 1$ threshold parameters for each y^r . As x_{kt} increases, so does the probability of observing larger-numbered outcomes m on the manifest y_{kt}^r . The only observed values in equation 1 are the y_{kt}^r on the left-hand side: all other parameters are estimated by the model.

To estimate x_{kt} and other auxiliary parameters of interest, we adopt a fully Bayesian approach. Theoretical considerations suggest that x is naturally bounded by a logical minimum and maximum value. Since x is a conceptual variable, we arbitrarily place its lower bound at zero (“none” of the latent characteristic) and its upper bound at one (“all” of the latent characteristic). We achieve a smooth trend in our estimate of the latent variable using a Bayesian random walk process (e.g., Martin and Quinn, 2002). Within country k , we assume that the latent value in year t has a Normal, but bounded, prior distribution that

is centered at the previous value of the latent variable in year $t - 1$;

$$x_{kt} \sim N(x_{k(t-1)}, \sigma_k^2) \mathcal{I}(0, 1). \quad (2)$$

The notation $\mathcal{I}(0, 1)$ indicates that x_{kt} can not exceed the unit interval. For year $t = 1$ we assume a non-informative Normal prior, also censored beyond zero and one. The variance parameters σ_k^2 , which are estimated separately for each country, capture the amount of temporal variation in x_{kt} . In countries where x_{kt} is relatively unchanged from year-to-year—typically due to countries remaining at the maximum or minimum level of x for the entire period of observation—values of σ_k^2 will be close to zero. In countries where x_{kt} experiences more substantial or rapid yearly changes, σ_k^2 can be larger. Letting σ_k^2 vary by country ensures that countries where x_{kt} varies greatly are not “oversmoothed” by comparison to countries where x_{kt} is more stable. The σ_k are each assigned vague uniform priors.

The Bayesian specification allows us to seamlessly handle the frequent occurrence of missing data (Jackman, 2000). In each country-year, up to R manifest ratings y_{kt}^r are observed. Country-years with greater numbers of observed y_{kt}^r will have more information from which to update x_{kt} , based on the estimated β_r and γ_k . When many y_{kt}^r are missing, the posterior estimate of x_{kt} will be closer to its prior distribution. In cases where *every* y^r is missing for a given country-year, the random walk process bridges the gap by connecting estimates of x_{kt} from the last year in which any y^r were observed to those in the next year with an observed value of y^r .

To complete the model specification, we place Normal prior distributions over the lowest estimated thresholds τ_{r1} for each manifest indicator, with mean -1 and variance 1. We ensure that the thresholds are strictly increasing over m by letting $\tau_{r2} = \tau_{r1} + \theta_{r1}$, $\tau_{r3} = \tau_{r2} + \theta_{r2}$, $\tau_{r4} = \tau_{r3} + \theta_{r3}$, and so forth, where $\theta_{rm} > 0$. The θ_{rm} are estimated, so also require priors; we assume standard Normal distributions censored at zero. Finally, we specify moderately informative priors over the β_r and γ_k parameters. When there are relatively few manifest

variables, as in most comparative research, sensibly chosen prior distributions can prevent coefficient estimates from increasing indefinitely (Bailey, 2001). We assign each β_r and γ_k a Normal prior distribution with mean 1 and variance 0.05, and again censor the distribution at 0 to ensure that both coefficient vectors are strictly positive. We find that any priors more diffuse than this may allow estimates of β_r and γ_k to grow unrealistically large.

Measuring Judicial Independence

We apply our model to the context of cross-national measures of judicial independence. Over the past decade, multiple teams of policy analysts have designed indicators of judicial independence on a global scale (Haggard, MacIntyre and Tiede, 2008). The trend is not surprising. Scholars generally agree that unconstrained government encourages corruption (Alt and Lassen, 2003), retards economic growth (Barro, 1997), increases the risk of violent conflict (Bueno de Mesquita et al., 2003) and regime instability (North, Summerhill and Weingast, 2000), and undermines the protection of human rights (La Porta et al., 2004). There is also a growing consensus that courts empowered to evaluate the legality (at least) of governmental actions, and which are capable of inducing compliance with their decisions, are critical components of constrained governance. For this reason, the international community spends considerable resources each year promoting judicial reform and tracking its success (Carothers, 2006). Just as many analysts have looked to track judicial independence and evaluate its effects, others have sought to explain it, and there is no shortage of theoretical arguments.² Although the majority of empirical studies have been conducted within particular countries, scholars are increasingly testing implications of these models in a cross-national setting (Ginsburg, 2003; Gibler and Randazzo, 2011; Hayo and Voigt, 2007). Thus, there are good reasons for the proliferation of indicators.

²For a review of these arguments in international and domestic politics, see Staton and Moore (2011).

A new approach

There is considerable evidence suggesting that the teams are generally on the right track (Ríos-Figueroa and Staton, 2010), and so naturally scholars are inclined to ask which indicator in particular serves their project best. Rather than attempt to find the perfect indicator, however, we believe that there are compelling reasons, both conceptual and pragmatic, to ask how we might draw on all of them. The paramount justification lies in the nature of the concept. To see how so, it is critical to define what we mean by “judicial independence.” First, we have in mind a behavioral or *de facto* concept. We are not considering institutional arrangements like appointment, removal or budgetary institutions, which might render judges more or less capable of promoting constrained government.³ Instead, we wish to say something about how judges actually carry out their duties. We consider a judge “independent” in so far as her decisions reflect her evaluation of the legal regard and in so far as those decisions are respected by government officials who disagree with them. A judiciary’s independence increases in the independence its judges. This definition reflects a common approach,⁴ but the critical point is this. Whether we restrict our definition of independence to autonomous decision-making as in the first part of our concept, or perhaps conceive of an independent judge as “impartial,” or “bound by the law” or qualify the matter in some other reasonable way, we must confront the fact that the concept we are looking to measure is *latent*. We simply cannot observe independence directly; we have to infer it.

The immediate consequence of this inferential challenge is that existing estimates of judicial independence are subject to error—they must be. And error likely emerges in a number of ways and at a number of stages. Research teams may differ slightly in the concept they seek to measure. Or, it may be that teams share a concept of independence yet nevertheless develop an indicator that combines independence with some other concept

³These rules may be important in their own right, and they present their own measurement challenges. However, they are directly observable.

⁴Specifically, it draws on the Cameron (2002) power concept of judicial independence. Indeed, we are comfortable with calling this concept “judicial power,” as well. See Ríos-Figueroa and Staton (2010) for a general discussion of concepts in existing studies.

(as in common measures of the “rule of law”). Even if teams share a concept and seek to measure it and only it, the precise operationalization of that concept can differ. And obviously, the raters these teams employ are unlikely to conduct the coding errorlessly. This is especially true if we believe that the difficulty of measuring judicial independence varies across countries and/or years.

There are a number of practical reasons for the approach we advocate, as well. Extant *de facto* indicators are positively correlated with each other and associated with outcomes anticipated by theories of judicial independence (e.g. political rights, economic development) (Ríos-Figueroa and Staton, 2010). That said, the indicators are more strongly related to each other among developed states than among underdeveloped states. Also, there are two significant problems of missing data. The first, as Table 1 suggests, is that the indicators span different ranges of years, with uneven overlap. The second is that even in years when research teams attempted to provide scores, there is still considerable missingness; and, this missingness is correlated with development. The direct consequence of not addressing these missing data concerns is to inflate confidence in the robustness of our studies to alternative indicators, since we are conducting our analyses on the set of states for which our indicators are more likely to agree. Our model makes use of the general agreement among the indicators, yet addresses concerns resulting from measurement error and missing data.

Data

We make use of eight manifest indicators, described in Table 1, that have been used to capture *de facto* judicial independence, and which are reviewed by Ríos-Figueroa and Staton (2010). The authors suggest that five are designed precisely to capture the concept we have in mind (p. 28). For two of them, Fraser and XCONST, it was unclear which concept of judicial independence was being measured. It also appears that the Howard-Carey indicator is designed to capture “autonomy.” Since an independent court under our concept must be autonomous, Howard-Carey should provide relevant information.

Variable	Measurement level	Years available	Percent missing	Source
Tate-Keith	ordinal; 3 categories	1990-2004	66%	Tate and Keith (2009)
Howard-Carey	ordinal; 3 categories	1992-1999	83%	Howard and Carey (2004)
CIRI	ordinal; 3 categories	1981-2009	38%	Cingranelli and Richards (2010)
XCONST	ordinal; 7 categories	1960-2008	19%	Marshall and Jagers (2010)
CIM	interval; 0–1	1960-2000	37%	Clague et al. (1999)
Feld-Voigt	interval; 0–1	1980-2003	77%	Feld and Voigt (2003)
PRS	interval; 0–6	1984-2008	60%	Ríos-Figueroa and Staton (2010)
Fraser	interval; 0–10	1995, 2000-2005	93%	Ríos-Figueroa and Staton (2010)

Table 1: *Eight variables used to scale latent judicial independence, and their availability.*

Tate-Keith, Howard-Carey, and CIRI are all drawn from the U.S. State Department Human Rights Country Reports. PRS, Feld-Voigt, and Fraser draw on international surveys of country experts. Polity IV’s measure of executive constraints, XCONST, is coded by the project’s team of experts. Finally, the Contract Intensive Money score (CIM) reflects the proportion of money that is held in banking institutions.⁵ The logic of this proxy measure is that individuals are more likely to keep their financial assets in banks when they believe that a state’s institutions designed to protect property rights are credible. The judiciary is central among institutions designed to do so.

Tate-Keith, Howard-Carey, CIRI, and Feld-Voigt each explicitly attempt to provide a score for the independence of a state’s judiciary.⁶ In contrast, XCONST, Fraser, and PRS are hybrid measures, reflecting a team’s evaluation of the judiciary along with other features of the legal system. For example, XCONST is designed to broadly reflect “constraints” on government authority. PRS is a measure of law and *order*. Fraser provides information on the rule of law, impartiality and the protection of property rights, as well as the independence of the judiciary. On their own, each of the hybrid measures might be less reliable indicators of the underlying concept, however, as the underlying concept is latent and very much related to the non-judicial independence features of these measures, their inclusion is highly reasonable.

⁵Specifically, CIM is “the ratio of non-currency money to the total money supply, or $(M2-C)/M2$ where $M2$ is a broad definition of money supply and C is currency held outside of banks,” (Clague et al., 1999, 188)

⁶Feld-Voigt is recorded as a single value that we assume to have held constant from 1980 to 2003.

We investigate 200 countries over the 50-year interval from 1960 to 2009. Because many countries were not in existence for the entire study period, there are a total of 8,197 potentially observable country-years in the data set. Between the eight indicators, data coverage is most consistent in the mid-1990s; and at least five variables are available in each year between 1984 and 2004. If every manifest variable had been measured in every country-year, there would be a total of 65,576 observed values of y_{kt}^r . In actuality, we only observe 26,939, for an overall missingness rate of 59%. This missingness is distributed unevenly across the eight indicators, with Fraser missing fully 93% of the possible country-years, and XCONST observed for all but 19%.

The first four indicators were coded as ordinal level variables by their original authors. We convert CIM into an eight-category measure with the first category for values lower than 0.3. The remaining seven categories bin observations into increments of 0.1 on the original scale, up to 1. This variable is highly left-skewed, as only 1.6% of observed values fall into the first group. Another 2.7% are in category two, and 4.6% are in category three. Feld-Voigt and Fraser we divide into six categories of equal width from their minimum to maximum value. PRS is already nearly categorical (most values are integers from 1 to 6), so to generate an ordinal measure, we round each rating up to the nearest whole number.

We implement the model specified by equations 1 and 2, with prior distributions as described, using the WinBUGS and R software packages (Lunn et al., 2000; Sturtz, Ligges and Gelman, 2005; R Development Core Team, 2011). Parameters x_{kt} , β_r , γ_k , τ_{rm} , and σ_k^2 are estimated using a Markov chain Monte Carlo sampling procedure. We produce three parallel chains of 1000 samples from the joint posterior distribution, discarding the first half as a lengthy burn-in period to avoid dependence on the start values. Inferences are based on the remaining 1500 simulated draws. Convergence is assessed by visual inspection of the three chains for adequate mixing (Cowles and Carlin, 1996; Gelman et al., 2004).

Results

The model produces estimated levels of judicial independence for every country and year in our data set. An electronic file containing all 8,197 of these values is available from the authors by request. As an initial validation of our results, we plot the estimates for the most recent year, 2009, along with an associated measure of uncertainty (Figure 1). The ordering predictably places countries such as North Korea, Cuba, and Libya at the low end of the scale, and countries like Finland, Australia, and Japan at the top.

The countries that are estimated to have the highest and lowest levels of judicial independence are also estimated with the *smallest* amount of posterior uncertainty. This is just as we should expect: the countries that are more difficult to rank are those towards the center of the scale. However, this result is precisely the opposite of what is found by standard IRT models that assume that the latent variable is unbounded. In the analysis of Pemstein, Meserve and Melton (2010), for example, the level of democracy in countries at the absolute top and bottom of the scale are estimated with the *greatest* amount of uncertainty. This is attributable to the “truncation inherent in the individual component scales” (p. 440). Once countries reach the “limits” of the scales, there is no more information from which to distinguish one highly democratic country from another. But in our analysis, when countries consistently demonstrate features that reveal very low or very high levels of judicial independence, the estimator can reliably place those countries at exactly the most extreme position on the latent scale.

Temporal trends

Theories of judicial independence commonly make predictions about changes over time; however, the prevalence of missing data in the extant indicators has largely prohibited quantitative scholars from even simply describing temporal variation in the concept. On some theoretical accounts, major political events should cause an abrupt change in judicial in-

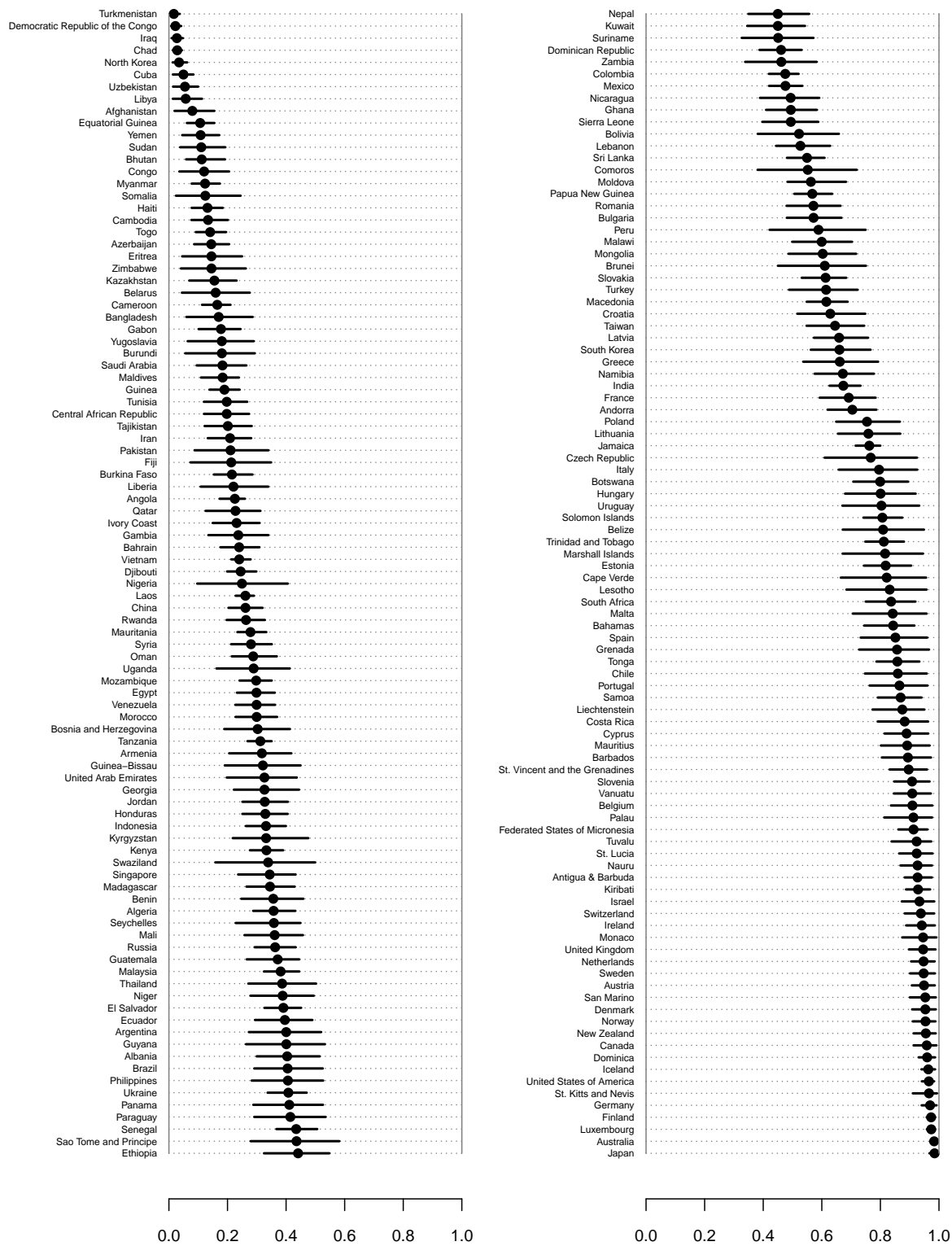


Figure 1: *Estimates of judicial independence in 191 countries in 2009. Error bars indicate 80% posterior credible intervals.*

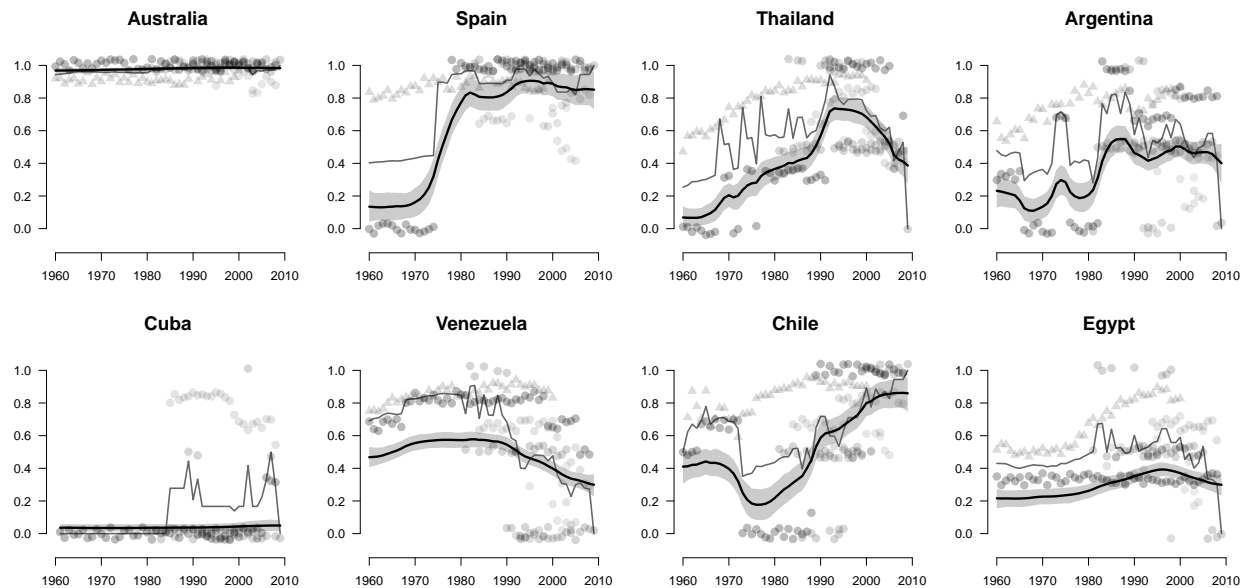


Figure 2: Trends in judicial independence in eight countries. Points denote the observed data, re-scaled to the unit range, and jittered for improved visibility. For indicators that were originally continuous, we plot the original values rather than the categorical y^r . More darkly shaded points indicate manifest variables with larger item discrimination, β_r . Seven of the eight y^r are shown as circles, but we plot CIM using triangles to make it stand out. The thick black line represents estimates of the latent x_{kt} , with an 80% posterior credible interval. The thinner, jagged line shows the result of re-scaling each of the original eight indicators to the unit range, and then taking the average of the observed values in each country-year.

dependence (e.g., North and Weingast, 1989; Ginsburg, 2003), whereas on other accounts, judicial independence should evolve more or less smoothly over time (e.g., Carrubba, 2009; Helmke, 2005). One of the most significant contributions of our model is its ability to estimate any number of different trends in the latent variable over time, based upon the highly flexible random walk prior process. The practical benefit to analysts is the possibility of conducting within-country, temporal comparisons across relatively long time intervals. Our measure is sensitive to both large changes in judicial independence, as well as more gradual or subtle shifts over time.

To illustrate, we select eight countries with a variety of temporal patterns of judicial independence, and plot the estimated x_{kt} , surrounded by an 80% credible interval (Figure 2).⁷ The observed data, re-scaled to the unit interval, are shown as points. More darkly shaded

⁷Plots showing the complete set of estimated time series are attached to the back of this paper.

points reflect indicators with larger item discrimination (β_r). We also compare our estimates to a naive descriptive estimate obtained by averaging together the available manifest indicators for each country-year. In each case, the assumptions of the model lend considerable statistical power to the estimator, filtering away a large amount of measurement error and stochastic noise, and allowing us to uncover small or higher-order trends with much greater reliability.

The model reveals a number of temporal patterns that should be familiar to judicial scholars. The first column highlights temporally invariant patterns that emerge at the top and bottom of the latent scale over the entire period of observation. For Australia, the observed indicators cluster tightly together at the top of the scale, reflecting the general consensus that the Australian judiciary, at least relative to the world, was considerably independent over the last 50 years. The Cuban series suggests precisely the opposite. As the time period covers the entire history of the Castro brothers' regime, it would have been surprising to estimate anything other than what we see in Figure 2. Yet, Cuba also allows us to consider a consequence of simply averaging the indicators. The light grey points near the top of the scale reflect the PRS measure, which picks up orderly societies as much as societies in which judiciaries are independent. This is, however, the only indicator that suggests anything other than low latent independence. When data are scarce, the mean will be overly sensitive to individual, discrepant indicators—here, spiking upwards as soon as PRS enters in 1984. Importantly, PRS exhibits relatively low discrimination and since it is in great disagreement with the remaining indicators, the model largely ignores its claims about Cuba. This is not to say that PRS is “wrong” about Cuba in an absolute sense. It may be quite right about Cuba’s law and order status. It does suggest that it does not do particularly well at revealing judicial independence.

It is more common for judicial independence to vary over time, in response to significant domestic political events and shifts in government policy. The second column displays cases in which there is a prolonged, unidirectional upward or downward trend. The change

is abrupt in Spain, responding very strongly to its transition to democracy in the period immediately following Franco’s death. This change likely reflects explicit efforts of reformers to change the nature of constitutional control in Spain via a system of centralized constitutional review (Guillen Lopez, 2007). But it also reflects the fact that the both Spanish legal culture and Spain’s judiciary was not entirely lined up with the Franco regime. There was clearly a piece of the Spanish judiciary willing to place limits on the state, if given the chance. Larkins (1996, 612) writes

In a survey of Spanish judges, political scholar José Toharia discovered that many were quite impartial, due to the fact that they held distinct values apart from the Franco government. For example, on four issues of extreme importance to the regime—the “law and order state” (including civil liberties), the legality of divorce, the use of languages other than Castillian Spanish, and the death penalty—Toharia’s data showed that a majority of Franco’s judges strongly disagreed with the government’s official positions... However, as Toharia proceeds to explain, this high level of impartiality was qualified by subtle limits on the insularity of judges and a significantly abridged scope of authority.

If these accounts are correct, then we should have observed an abrupt change in Spanish judicial independence upon the transition to democracy.

Venezuela reflects a different path. The Chavez period has been associated with a gradual erosion of judicial independence, through court packing at various levels and targeted purges (Taylor, 2009). This change is picked up by the latent measure. Critically, the average suggests a massive drop in 1989; however, all that has “happened” is that additional indicators become available at the bottom of the scale. The Venezuelan panel reveals another subtle feature of the model. Note that while the average tracks the two (and only) observed indicators during the 1960s and 1970s, the smoother is consistently lower. The reason is that observed values between 0.7 and 0.8 on the re-scaled XCONST and CIM measures are

associated with middling scores for the *other* indicators. For this reason, the model estimates independence in Venezuela to be closer to the center of the scale during this period.

The final set of examples shows states in which judicial independence has turned, sometimes multiple times—and demonstrates the capacity of the model to identify non-monotonic patterns in judicial independence. The Chilean series reflects constraints on the judiciary imposed by the Pinochet regime, which were largely removed after the transition (Scribner, 2011).⁸ The Thai panel mirrors this pattern. The 1997 Thai constitutional reform was designed to confirm democratic changes in the regime following the Black May Uprising of 1992, and gave new authority to the judiciary to investigate allegations of political corruption. The estimates reflect an increase in judicial independence beginning at the end of the 1980s and peaking in the 1990s. Ginsburg (2009) argues, however, that what gains might have been made in the period following the 1997 reform were undermined by Prime Minister Thaksin, who came to power in 2001. He writes, “Gradually, Thaksin began to influence all the independent political institutions, including the Constitutional Court and those designed to prevent corruption.” As in Cuba, using the simpler average trend produces a series of misleading peaks and valleys, due to the relative paucity of data prior to 1980.

The model can also distinguish more subtle trends in judicial independence, as in Egypt. The Egyptian series suggests a rise beginning around 1980, followed by a fall starting in the late 1990s. About this period, Brown (2002, 151-152) writes, “After 1979 (especially after the mid-1980s when the new appointment procedure had begun to seriously affect the composition of the [Constitutional] Court), the Court rapidly distinguished itself as the boldest and most independent judicial actor in Arab history.” Yet in the late 1990s, “The presidency of the Supreme Constitutional Court fell vacant with the retirement of the activist ‘Awad al-Morr. the vacancy was used to pressure the Court into accepting a diminution in its authority to issue retroactive judgments.”

⁸But see Hilbink (2007) who argues that the Chilean judiciary’s independence was not compromised during this period.

Argentina’s panel is highly informative and speaks directly to the validity of the measure. We would expect a measure of governance in Argentina to be highly unstable: observing the peaks and valleys in Figure 2 is not surprising. And in light of theories like Helmke’s (2005), we would expect to observe changes in the series as regimes de-stabilize. It is particularly interesting that the estimate picks up an upward change in judicial independence beginning in 1980, three years prior to the fall of the junta. This is consistent with Helmke’s argument, but is also important to note that the measure responds to features of Argentine judicial politics that are independent of regime change. Chávez, Ferejohn and Weingast (2011) argue that patterns of judicial independence in Argentina have tracked the fragmentation of government closely, precisely because it was difficult to discipline the court absent inter-party coordination. Specifically, they argue that since government was divided during the Alfonsín era (1983-1989), the judiciary found space to “challenge the executive” (p. 237). Yet, during the Menem period (1989-1997), where government was unified, judicial independence was deeply compromised. Our estimates support these claims. Indeed, there is a pronounced increase in judicial independence beginning in the early 1980s, followed by an abrupt change at the end of the decade, precisely when the President began to pack the Supreme Court.

Assessing the indicators

The model returns information not only about latent judicial independence, but the indicators themselves. Indeed, inferences about the latent variable depend upon the relationship between x and each of the manifest y^r in country k . This is captured in our model by the discrimination parameters β_r , threshold values τ_{rm} , and country-reliability parameters γ_k . Estimates of β_r range from 2.46 for Feld-Voigt to 8.16 for XCONST, and are characterized by extremely minimal levels of uncertainty, with posterior standard deviations in the range of approximately 0.1 to 0.15. As is apparent in Figure 2, indicators with larger discrimination parameters exhibit greater “pull” on the latent variable.

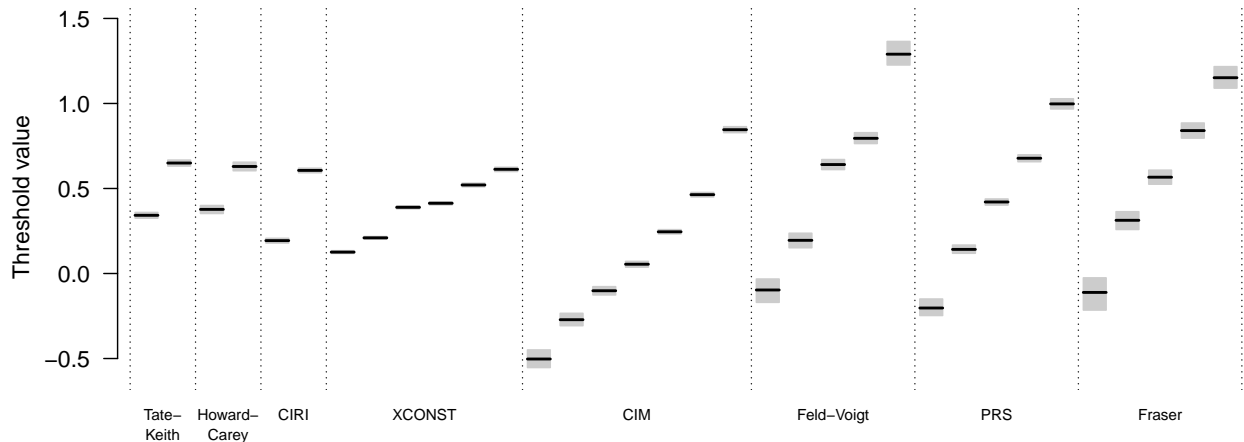


Figure 3: *Comparison of threshold estimates τ_{rm} for eight indicators of judicial independence. Shading denotes areas of 95% highest posterior density.*

But the relative positioning of the threshold estimates for each indicator is also extremely important—especially for scaling x_{kt} during intervals when data are sparse. In effect, the thresholds describe how outcomes on each variable “align” with one another along the latent scale. Different indicators are better at distinguishing values of x_{kt} at different levels of the latent variable (Figure 3). The three-category indicators Tate-Keith, Howard-Carey, and, to a lesser extent, CIRI, all partition the scale in a relatively similar fashion: country-years rated 1 (below the lowest threshold) on one variable are likely to be rated 1 on the others as well. By comparison, ratings of 1 on Tate-Keith and Howard-Carey roughly correspond to ratings of 3 or less on XCONST; and, continuing to read across Figure 3, ratings of up to 5 or 6 on CIM. Because CIM is so left-skewed, a country-year can have a relatively high observed value (say, 5 out of 8) and still belong at the low end of the latent scale. The consequences of this are immediately apparent in Figure 4. CIM, which is plotted using triangular points, is consistently above the smoothed latent trend. While the trend line for the mean is fooled by these “large” values of CIM, our model recognizes that even when CIM is as high as 0.7 (corresponding to category 5), judicial independence should still be considered low. Similarly, when CIM is above 0.9 (category 8), the country-year is very

Ten lowest	γ_k	Ten highest	γ_k
Swaziland	0.39 (0.32, 0.46)	Cyprus	2.07 (1.88, 2.26)
Lesotho	0.56 (0.45, 0.67)	Cameroon	1.98 (1.82, 2.15)
Qatar	0.70 (0.61, 0.79)	Sweden	1.96 (1.76, 2.16)
Kuwait	0.75 (0.67, 0.83)	Tanzania	1.95 (1.80, 2.11)
Georgia	0.75 (0.63, 0.87)	Poland	1.93 (1.74, 2.11)
Bahrain	0.76 (0.68, 0.85)	Portugal	1.90 (1.73, 2.08)
Central African Republic	0.83 (0.73, 0.93)	Ivory Coast	1.89 (1.72, 2.05)
Czech Republic	0.93 (0.77, 1.11)	United States of America	1.89 (1.69, 2.09)
Bhutan	0.96 (0.81, 1.11)	Ireland	1.88 (1.68, 2.08)
Myanmar	0.97 (0.87, 1.09)	Mozambique	1.88 (1.71, 2.05)

Table 2: *Ranking the ten best- and worst- measured countries, by estimated γ_k . Parentheses indicate 80% credible intervals.*

likely to be scaled as close to 1. Country-years manifesting the highest-category outcomes on Feld-Voigt, PRS, and Fraser are even more likely to appear at the top of the latent scale.

Estimates of the country-level γ_k parameters suggest that some countries are indeed less conducive than others to scaling judicial independence based on the chosen y^r . Following equation 1, larger values of γ_k amplify the effects of all eight β_r in country k , leading the trend in x_{kt} to respond more strongly to the observed data. Where γ_k is smaller, the data are less informative about the latent value and as a result, the prior receives greater weight and the trend changes more gradually over time. In Table 2, we list the ten countries with the highest and lowest estimated γ_k . At the low end, countries such as Georgia, Kuwait, and the Central African Republic are especially poorly measured: the manifest indicators do not trend together in a manner that is systematically similar to other countries in the sample. At the high end, developed countries such as the USA, Sweden, and Ireland appear alongside the African states of Cameroon, Tanzania, and Mozambique.

When γ_k is as high as 2, the product $\beta_r \gamma_k$ in equation 1 can generate coefficients of 15 or more on certain items. It may nevertheless be the case that judicial independence in country k remains relatively consistent over the entire study period, as happens in the United States, for example. Large values of γ_k only create the potential for substantively meaningful effects on x_{kt} , in countries where the latent trend actually varies over time. A more consistent feature of γ_k is its association with the size of the posterior credible interval surrounding

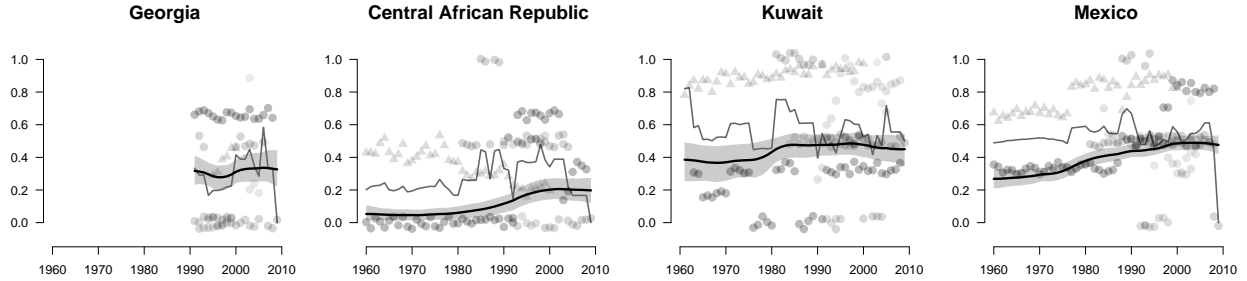


Figure 4: Comparing countries by estimated γ_k : low values in Georgia (0.75), Kuwait (0.75), and the Central African Republic (0.83); and a much higher value in Mexico (1.83).

estimates of x_{kt} : smaller γ_k reveal greater posterior uncertainty in the value of the latent variable (Figure 4). In Kuwait, for example, a variety of different indicators simultaneously rate the country at both the maximum and minimum of the scale for nearly the entire study period. The result is a nearly flat, and fairly uncertain, trend. We compare this to Mexico, which with $\gamma_k = 1.83$ is among the “best-measured” countries in our sample. It is reasonable that both measurement quality and conceptual clarity would be high for a country adjacent to the United States, about which much is known. The trend in judicial independence in Mexico is nevertheless relatively flat over the observed period.

From a theoretical perspective, countries with small γ_k are interesting for other substantive and diagnostic reasons. These estimates can direct researchers to the countries or regions that may require greater attention to be paid to reducing measurement error. Alternatively, they may indicate countries where existing theories about the consequences of the latent variable are failing to apply, helping to guide the development of new theories that generalize more broadly around the world. Many of the most poorly measured countries in our analysis, for example, are in the Middle East and central Asia. Why this should be the case is not immediately apparent, but warrants a closer examination.

Conclusion

Latent concepts pervade comparative politics and international relations. The common approach to measuring them has involved the use of single proxy instruments, where scholars at best consider the robustness of findings to alternative proxies. Researchers have recently approached the problem through the application of measurement models designed for the task. In this paper, we develop a heteroscedastic graded response IRT model designed for the kind of time series, cross-sectional data common to studies of comparative politics and international relations, and apply it to the challenge of measuring latent judicial independence.

We stress two features of our approach which represent large substantive improvements over existing latent variable models. The first involves the assumption that latent judicial independence follows a Bayesian random walk prior process. This permits us to smooth our estimates over time. As many IR and CP concepts measured in TSCS data are relatively slow-moving over time, this assumption gives us additional statistical power. In a context where different indicators generally agree but measurement error is significant, smoothing allows us to cut through considerable noise. Second, by bounding the latent variable, we obtain uncertainty estimates that make sense. We should be more confident about the placement of states at either end of the scale than we are about states in the middle. Bounding the latent variable produces this effect.

There are a number of implications for the study of judicial independence. Clearly, comparative judicial scholars have been limited by extant indicators of the concept. Related patterns of measurement error and missing data have not only complicated analyses, but they have rendered some kinds of studies simply impossible to conduct with anything but a rough proxy. The sheer increase in data that we provide addresses these practical problems. We are now in a position to trace judicial independence systematically over a relatively long time period. Importantly, the estimates are neither too sensitive to severe changes in one or two indicators nor are they completely insensitive to massive changes in political context.

We hope that this feature of the measure will allow scholars to more precisely evaluate claims about changes over time, claims that suggest both dramatic and subtle change.

It is also transparent that no theory of judicial independence that anticipates only one kind of development over time can explain what we observe. Trends around the globe simply fit multiple patterns. Judicial scholars have always known this, but our estimates confirm the point clearly. Some states are highly stable. Some are highly volatile. Others exhibit change, but in one direction, while still others experience considerable backslides or single changes “for the better.” We hope that our measure will contribute to the further development of theories that are conditional and sensitive to context.

Another nice feature of the model is its ability to speak to the difficulty of measuring particular countries. We hope that our findings on this score will influence how research teams evaluate their measurement efforts, perhaps by re-thinking why we seem to get some—but not all—countries consistently correct. The findings also offer a tool for funders to evaluate whether to encourage particular work in the states where we seem to be doing a relatively mediocre job.

Finally, we believe it is important to stress that scholars who choose to neither use our measure, nor estimate their own, will do well to at least consider averaging the scores to which they have access. Figure 2 certainly suggests that an average is far more unstable than the estimate we provide, and there are years where it seems artificially high or low. That said, it does a reasonably good job of aggregating the scores. In a pinch, averaging these series is better than selecting one indicator on more or less arbitrary grounds.

References

- Alt, James E. and David Dreyer Lassen. 2003. "The political economy of institutions and corruption in American states." *Journal of Theoretical Politics* 15(3):341.
- Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(2):215–232.
- Arel-Bundock, Vincent and Walter R. Mebane, Jr. 2011. "Measurement Error, Missing Values and Latent Structure in Governance Indicators." Paper presented at the Annual Meeting of the American Political Science Association, Seattle, WA.
- Bailey, Michael. 2001. "Ideal Point Estimation with a Small Number of Votes: A Random-Effects Approach." *Political Analysis* 9(3):192–210.
- Barro, Robert J. 1997. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge: MIT Press.
- Brown, Nathan J. 2002. *Constitutions in a nonconstitutional world: Arab basic laws and the prospects for accountable government*. State University of New York Press.
- Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson and James Morrow. 2003. *The Logic of Political Survival*. MIT Press.
- Cameron, Charles M. 2002. Judicial Independence: How Can You Tell It When You See it? And, Who Cares? In *Judicial Independence at the Crossroads: An Interdisciplinary Approach*, ed. Steven B. Burbank and Barry Friedman. New York: Sage Publications Inc. pp. 134–147.
- Carothers, Thomas. 2006. *Promoting the rule of law abroad: in search of knowledge*. Carnegie Endowment for Intl Peace.
- Carrubba, Clifford J. 2009. "A Model of the Endogenous Development of Judicial Institutions in Federal and International Systems." *Journal of Politics* 71(1):1–15.
- Chávez, Rebecca Bill, John A. Ferejohn and Barry R. Weingast. 2011. A Theory of the Politically Independent Judiciary: A Comparative Study of the United States and Argentina. In *Courts in Latin America*. New York: Cambridge University Press.
- Cingranelli, David L. and David L. Richards. 2010. "The Cingranelli Richards (CIRI) Human Rights Database Coding Manual." available online at: <http://ciri.binghamton.edu/documentation.asp>.
- Clague, Christopher, Philip Keefer, Stephen Knack and Mancur Olson. 1999. "Contract-Intensive Money: Contract Enforcement, Property Rights, and Economic Performance." *Journal of Economic Growth* 4(2):185–211.

- Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Legislative Behavior: A Unified Approach." *American Political Science Review* 98(2):355–370.
- Cowles, Mary Kathryn and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91(434):883–904.
- Feld, Lars P. and Stefan Voigt. 2003. "Economic Growth and Judicial Independence: Cross-country Evidence Using a New Set of Indicators." *European Journal of Political Economy* 19(3):497–527.
- Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.
- Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis; Second Edition*. Chapman & Hall/CRC.
- Gibler, D.M. and K.A. Randazzo. 2011. "Testing the Effects of Independent Judiciaries on the Likelihood of Democratic Backsliding." *American Journal of Political Science* 55(3):696–709.
- Ginsburg, Tom. 2003. *Judicial Review in New Democracies: Constitutional Courts in Asian Cases*. New York: Cambridge University Press.
- Ginsburg, Tom. 2009. "Constitutional afterlife: The continuing impact of Thailand's post-political constitution." *International journal of constitutional law* 7(1):83.
- Guillen Lopez, E. 2007. "Judicial Review in Spain: The Constitutional Court." *Loyola of Los Angeles Law Review* 41:529.
- Haggard, Stephan, Andrew MacIntyre and Lydia Tiede. 2008. "The Rule of Law and Economic Development." *Annual Review of Political Science* 11:205–234.
- Hayo, Bernd. and Stefan Voigt. 2007. "Explaining de facto judicial independence." *International Review of Law and Economics* 27(3):269–290.
- Helmke, Gretchen. 2005. *Courts under Constraints*. Cambridge: Cambridge University Press.
- Hilbink, Lisa. 2007. *Judges beyond politics in democracy and dictatorship: lessons from Chile*. Cambridge University Press.
- Honaker, James and Gary King. 2010. "What to do about Missing Values in Time Series Cross-Section Data." *AJPS* 54(3):561–581.
- Howard, Robert M. and Henry F. Carey. 2004. "Is an Independent Judiciary Necessary for Democracy?" *Judicature* 87(6):284–290.

- Jackman, Simon. 2000. “Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation.” *Political Analysis* 8(4):307–332.
- Johnson, Valen E. and James H. Albert. 1999. *Ordinal data modeling*. New York: Springer-Verlag.
- Jolliffe, I.T. 2002. *Principal Component Analysis, Second edition*. New York: Springer-Verlag.
- La Porta, Rafael, Florencio López de Silanes, Cristian Pop-Eleches and Andrei Shleifer. 2004. “Judicial Checks and Balances.” *Journal of Political Economy* 112(2):445–470.
- Larkins, Christopher M. 1996. “Judicial independence and democratization: A theoretical and conceptual analysis.” *The American Journal of Comparative Law* 44(4):605–626.
- Lauderdale, Benjamin E. 2010. “Unpredictable Voters in Ideal Point Estimation.” *Political Analysis* 18(2):151–171.
- Lunn, D.J., A. Thomas, N. Best and D. Spiegelhalter. 2000. “WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility.” *Statistics and Computing* 10:325–337.
- Marshall, Monty and Keith Jagers. 2010. “Polity IV Project: Political Regime Characteristics and Transitions, 1800–2004.”
- Martin, Andrew and Kevin Quinn. 2002. “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999.” *Political Analysis* 10:405–419.
- North, Douglass and Barry Weingast. 1989. “Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in 17th Century England.” *Journal of Economic History* 49(4):803–832.
- North, Douglass, William Summerhill and Barry Weingast. 2000. Order, Disorder, and Economic Change: Latin America vs. North America. In *Governing for Prosperity*, ed. Bruce Bueno de Mesquita. New Haven: Yale University Press.
- O’Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley and Tim Rakow. 2006. *Uncertain Judgments: Eliciting Experts’ Probabilities*. West Sussex: Wiley.
- Pemstein, Daniel, Stephen A. Meserve and James Melton. 2010. “Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type.” *Political Analysis* 18:426–449.
- Przeworski, Adam, Michael 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.
- Quinn, Kevin M. 2004. “Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses.” *Political Analysis* 12(4):338–353.

- R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. <http://www.R-project.org>.
- Ríos-Figueroa, Julio and Jeffrey K. Staton. 2010. “An Evaluation of Cross-National Measures of Judicial Independence.” Emory University.
- Scribner, Druscilla. 2011. Courts, Power, and Rights in Argentina and Chile. In *Courts in Latin America*. New York: Cambridge University Press.
- Staton, Jeffrey K. and Will H. Moore. 2011. “Judicial Power in Domestic and International Politics.” *International Organization* 65:553–87.
- Stokes, Susan C. 2001. *Mandates and Democracy: Neoliberalism by Surprise in Latin America*. New York: Cambridge University Press.
- Sturtz, Sibylle, Uwe Ligges and Andrew Gelman. 2005. “R2WinBUGS: A Package for Running WinBUGS from R.” *Journal of Statistical Software* 12(3):1–16.
- Tate, Neal C. and Linda Camp Keith. 2009. “Conceptualizing and Operationalizing Judicial Independence Globally.” Working Paper.
- Taylor, Matthew M. 2009. “A Model of Judicial Independence with Illustration from Chavez’s Venezuela.” Paper presented at the annual American Political Science Association Meeting, Toronto, Canada.
- Treier, Shawn and Simon Jackman. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science* 52(1):201–217.
- van der Linden, Wim J. and Ronald K. Hambleton. 1997. *Handbook of modern item response theory*. New York: Springer.
- Voeten, E. 2007. “The Politics of International Judicial Appointments: Evidence from the European Court of Human Rights.” *International Organization* 61:669–701.

