Political Science Research and Methods Vol 2, No. 1, 1–31 April 2014

doi:10.1017/psrm.2013.15

Dynamic Patterns of Human Rights Practices*

KEITH E. SCHNAKENBERG AND CHRISTOPHER J. FARISS

The science of human rights requires valid comparisons of repression levels across time and space. Though extensive data collection efforts have made such comparisons possible in principle, statistical measures based on simple additive scales made them rare in practice. This article uses a dynamic measurement model that contrasts with current approaches by (1) accounting for the fact that human rights indicators vary in the level of information they provide about the latent level of repression, (2) allowing realistic descriptions of measurement uncertainty in the form of credible intervals and (3) providing a theoretical motivation for modeling temporal dependence in human rights levels. It presents several techniques, which demonstrate that the dynamic ordinal item-response theory model outperforms its static counterpart.

Did repression decrease in Guatemala after the end of the Cold War? Is Uzbekistan more repressive today than when it first emerged as an internationally recognized state? Was Latin America a more repressive region than the rest of the world in the 1980s? The promise of a science of international human rights relies on the ability to provide transparent and realistic answers to descriptive questions like these. Unfortunately, as we argue in this article, the measures currently used in the quantitative human rights literature are ill suited to the task of comparing levels of repression between countries or over time. In this article, we provide a theoretically motivated measurement model that enhances researchers' ability to compare repression levels across time and space.

This project provides several distinct contributions to political science research. We emphasize the special importance of measurement in human rights research relative to many other areas of political science. Precise scoring of repression levels using available information in different countries is directly useful to international and non-governmental organizations (NGOs). Moreover, considerable scholarly attention focuses on the efficacy of "naming and shaming" campaigns, which are claimed to improve human rights practices.[1] If these claims are true, the availability of messages that can more precisely

* Keith Schnakenberg is a graduate student in the Department of Political Science at Washington University, St. Louis (keschnak@wustl.edu). Christopher Fariss is an Assistant Professor in the Department of Political Science, Pennsylvania State University (cjf0006@gmail.com). We thank Chad Clay, Jesse Driscoll, James Fowler, Jeff Gill, Miles Kahler, David Lake, Yon Lupu, Will Moore, Amanda Murdie, David Richards, Guillermo Rosas and Sebastian Saiegh for helpful comments. Earlier versions of this paper were presented at the 2012 meeting of the Midwestern Political Science Association, the 2012 meeting of the International Studies Association, the 2011 meeting of the American Political Science Association and the 2010 meeting of the Southern Political Science Association. We would also like to acknowledge the Southern Political Science Association for travel grants. The data and code used in this paper and the latent variable estimates from the dynamic models are publicly available at http://dvn.iq.harvard.edu/dvn/dv/HumanRightsScores. Both authors contributed to the study design, data collection, analysis and preparation of the manuscript.

[1] For example, Keck and Sikkink 1998; Meernik *et al.* 2012; Murdie and Davis 2012.

discriminate between repression levels in different countries may enhance the persuasive power of such techniques.

The model developed in this article allows descriptive human rights questions like those posed above to be answered in the form of a probability. The resulting model estimates provide the first source of information that allows for direct, probabilistic comparisons of levels of human rights abuse between cases. Until now, analysts who wished to compare two cases had to rely on case-specific evidence or additive human rights indices. Case-specific evidence limits the number of comparisons that can be made and does not provide an estimate that can be compared alongside other pairwise comparisons. The additive indices that predominate the quantitative human rights literature can provide yes or no answers, but rely on unrealistic assumptions about the data-generating process and do not allow uncertainty estimates. For example, two countries that receive the same score on one of the standard additive human rights scales are assumed to be the same with probability 1, and if another country-year pair receives different scores, then they are assumed to be different with probability 1. Furthermore, the human rights estimates generated by our model are interval-level measures, which simplifies data analysis for applied human rights researchers.[2]

The article also provides a methodological contribution to international relations scholars more generally. Though ordinal measurement models are currently used elsewhere in international relations (Treier and Jackman 2008; Pemstein, Meserve and Melton 2010), these models assume that the observed indicators are independent conditional on the value of the trait to be estimated. We demonstrate, however, that this is an overly strong assumption in the case of human rights data. In contrast to this approach, we use hierarchical priors that allow the estimated latent respect for human rights in a country in a particular year to depend on that country's value in the previous year.

Though we justify the dynamic measurement model for estimating latent human rights theoretically, we believe that a dynamic model similar to ours would be useful for measuring democracy and other temporally dependent processes. Our results suggest that our dynamic approach should be incorporated into models of democracy and other concepts used in time-series cross-sectional studies. Finally, we present three model comparison statistics—deviance information criterion, posterior predictive checks and predictive validity tests—which all demonstrate that the physical integrity and empowerment estimates produced by the dynamic model outperform those produced by a static model similar to the one used by Pemstein, Meserve and Melton (2010) and Treier and Jackman (2008).

This article proceeds as follows. In Section 2 we review existing methodological approaches for modeling latent variables and measuring uncertainty and introduce the Bayesian ordinal item-response theory (O-IRT) model and the theoretical justification for our proposed dynamic version(DO-IRT). Section 3 applies the DO-IRT and O-IRT models to the Cingranelli and Richards (CIRI) physical integrity rights and empowerment rights data. Section 4 presents parameter estimates obtained from the DO-IRT and O-IRT models for both physical integrity and empowerment rights and several probabilistic answers to the descriptive questions posed above. Section 5 discusses several model fit

[2] Some previous approaches in the human rights literature (Landman and Carvalho 2009) also produce interval-level variables from a simple factor analysis, which provides the authors with easy access to improved methods for dealing with the features of time-series cross-sectional data. As we discuss, our model uses assumptions that are more appropriate for the ordinal nature of the data.

statistics and demonstrates the improvement of the dynamic model over the static version by conducting posterior predictive checks and comparing the latent variables to the existing additive scales. Section 6 demonstrates how to incorporate the new measures into applied research. Finally, Section 7 concludes.

MEASUREMENT OF HUMAN RIGHTS PRACTICES

There is considerable debate in the human rights and repression literatures regarding the appropriate ways to measure relevant concepts.[3] Scholars have recently taken issue with the CIRI data because of its use of additive scales:

> Specifically, we question the logic of summing these categories to establish this picture because, in doing so, users must make the assumption that an act of torture is equivalent to a disappearance or that an extrajudicial killing is equivalent to an instance of arbitrary imprisonment (Wood and Gibney, 2010).

Though this has sometimes been interpreted as a criticism of the data source, we note that this objection applies not to the source of the data but to the additive scale as a model for aggregating indicators. The modeling approach used in this article helps alleviate existing concerns with the kinds of data that are used to provide quantitative answers to human rights questions. The criticism made by Clark and Sikkink (forthcoming) is directed at the documentary sources of the data and indirectly at the overreliance on the data derived from these sources. Though we are not able to confront this issue directly in this article, we do make some suggestions about extending our model, which can address the criticism of the data sources, given the availability of other sources of information.

Although we ultimately agree with Wood and Gibney that the assumption of equal weights between indicators is inadequate, the importance of utilizing multiple indicators of the same concept should not be underestimated. As Jackman (2008) points out, a researcher with only one indicator of a latent construct is unable to determine how much variation in the indicator is due to measurement error as opposed to variation in the latent construct.

Furthermore, the quality of the inferences made about repression levels depends on clearly specifying a theoretically informed model that best approximates reality, or in modeling terms, the data-generating process. Alison Brysk (1994) makes this point in an essay that highlights the difficulties inherent in the measurement of repressive actions. She notes "the importance of careful specification of the political processes being modeled through measurement and explicit justification of the use of particular measures to represent those processes" (1994, 692). We take up this challenge here and develop a model that explicitly assumes that repressive practices are related over time in that the pattern of abuses committed today might change the use of (or even the need for) future repressive actions for a certain period of time (Duvall and Stohl 1983; Stohl *et al.* 1986). Our model is the first in the human rights literature to formalize this idea in the

[3] See, for example, Brysk 1994; Donnelly and Howard 1988; Jabine and Claude 1992; Goodman and Jinks 2003; Lopez and Stohl 1992; Landman 2005; Landman and Carvalho 2009; McCormick and Mitchell 1997; Poe 1990, 1991, 2004; Poe, Carey and Vazquez 2001 and most recently Clark and Sikkink (forthcoming). Interested readers should also consult earlier papers from this debate, which are contained in Jabine and Claude (1992) and Claude and Jabine (1986).

measurement of repression. We then assess the usefulness of incorporating this idea into the model by comparing it to a model that does not make such an assumption.

Our model assumes there is an underlying trait that can be estimated using observed outcomes (that is, the items or responses). We are therefore able to focus on estimating a single physical integrity latent variable using the four physical integrity indicators measured by Cingranelli and Richards (2012a) and a single empowerment latent variable using seven empowerment indicators, which we describe below. This distinction is important, because Landman and Carvalho (2009) include measures that are not physical integrity rights in their factor analysis. However, these variables may be better captured by the Cingranelli and Richards (2012a) Empowerment Index. We therefore choose to build on the considerable research conducted by Cingranelli, Richards and their co-authors and estimate our dynamic latent variable model using only the indicators coded as part of the CIRI data project.[4]

To estimate the physical integrity and empowerment latent variables, we build on existing IRT models, which are increasingly important in political science. Static measurement models in political science are well developed for use with binary data, and have been developed especially for the case of recovering the ideal points of political actors. The seminal work on measuring ideal points from roll-call data is the NOMINATE project (Poole 2005; Poole and Rosenthal 1991, 1997), which employs a maximum likelihood approach to the problem and remains the most widely used model for this purpose in Congressional politics. This model has been applied in comparative politics (Desposato 2006) and to the United Nations General Assembly (Voeten 2000). The Bayesian approach to the problem, based on IRT, was used by Clinton, Jackman and Rivers (2004). Martin and Quinn (2002) extended the approach to incorporate dynamics and applied the model to the US Supreme Court. Political scientists have now published many extensions and applications of the binary item-response model (Aleman and Saiegh 2007; Lauderdale 2010; Rosas and Shomer 2008).

The item-response model related to the one employed by Clinton, Jackman and Rivers has been extended to include ordinal items.[5] The ordinal item-response model was applied to estimates of democracy by Treier and Jackman (2008). Quinn (2004) developed a model that incorporates both ordinal and interval responses and applies it to data on political risk. Pemstein, Meserve and Melton (2010), in the Unified Democracy Scores project, apply a model similar to Quinn's to aggregate several measures of democracy into a single scale. In this article, we build on the standard ordinal item-response model like the one employed by Treier and Jackman (2008).

Our methodological contribution is to extend ordinal item-response models into a dynamic setting. This endeavor has already been successful in the case of binary item-response models. The DW-NOMINATE procedure (Poole and Rosenthal 1997) is a dynamic version of W-NOMINATE which estimates idea points as a function of ideal points in the previous time period. Like the model presented in this article, the Martin and Quinn model accounts for temporal dependence in the data by specifying a prior for each value of the latent variable centered at the estimated latent variable from the same unit in

[4] Future research could combine additional indicators in a manner similar to Landman and Carvalho (2009). We leave this to future work in order to directly compare the static and dynamic latent variables generated by our model with the original additive indices created by Cingranelli and Richards (2012a).

[5] See Albert and Johnson (1999) for a detailed explanation of these models, which are also called "graded response models".

the previous time period. A few other approaches to dynamic measurement can be found in the literature. Rosas (2009), for example, estimates a latent trait model that places an autoregressive structure on the covariance between factor scores. To our knowledge, this article provides the first dynamic ordinal IRT model.

A MEASUREMENT MODEL OF HUMAN RIGHTS RESPECT

The strategy used in this article is to derive and then apply a measurement model to the CIRI human rights dataset. This data is used to construct two additive scales that are commonly used in the quantitative human rights literature. Though a strength of our model is that it can easily be extended to incorporate information from a wide variety of sources, the CIRI dataset is ideal for introducing the method because the reliabilities of the two CIRI scales are already established and because they facilitate comparison between our measurement model and the additive approach.[6] By reliable, we mean that the CIRI variables consistently represent the content of the human rights reports published annually by the US State Department and Amnesty International, based on their own coding criteria.

Though the CIRI scales are reliable, some scholars question their validity.[7] Wood and Gibney (2010) question the precision of the CIRI coding guidelines to categorize the reports' content. Clark and Sikkink (forthcoming) question the CIRI dataset's ability to compare coded reports from earlier periods that were quantifiably less informative than more recent ones based on raw word counts taken directly from some of the US State Department reports. The argument made by Wood and Gibney (and more recently by Clark and Sikkink) relates to the validity of the CIRI scales relative to the theoretical construct of interest: respect for human rights. If Cingranelli and Richards were only interested in accurately measuring the content of the reports, then Clark and Sikkink would have little reason to question the validity of the resulting variables, because Clark and Sikkink's argument is based on changes to the content of the human rights reports themselves. This is an important theoretical distinction that is often overlooked when the CIRI data and other variables based on the human rights reports are presented as measurements of repression rather than of *reported* repression.

Our model is based only on the CIRI data, which allows us to directly address the issue raised by Wood and Gibney (2010) but not Clark and Sikkink (forthcoming). However, the model is extendable and can incorporate new theoretical insights and data, which can then be used to assess the validity of the resulting estimates relative to the theoretical construct of interest: *respect for human rights*. Model comparison techniques are developed below that help us establish the validity of our dynamic measurement model relative to an alternative. Future research will be able to explore how these and other variables vary in the level of information they provide about repression by introducing additional information about repression and human rights into our dynamic measurement model and then using the same

[6] The average inter-coder reliability score for the CIRI variables is 0.944 (Cingranelli and Richards 2012a). Because these scores are all high, this information is unlikely to make a difference in the models we develop below. To the extent that this is an issue, however, it will be reflected in the item-discrimination parameters for the various variables. It would be useful to know which country-years generated coder disagreement in the estimation of these scores; however these data are not available. If we had the separate scores for the coders we would be able to use this information to further improve our proposed model.

[7] See, for example, Clark and Sikkink forthcoming; Wood and Gibney 2010.

model comparison techniques we describe below. In the next subsection, we introduce the data and the justification for our proposed dynamic measurement model (DO-IRT).

Human Rights Indicators

The Cingranelli and Richards (1999) Physical Integrity Scale is a single 9-point ordinal scale that ranges from 0 to 8 and is measured from country practices documented in human rights monitoring reports by the US State Department and Amnesty International. The CIRI data use this information to code violations of four individual physical integrity variables (political imprisonment, torture, extra-judical killing and disappearances).[8] As part of the larger CIRI data project Richards, Gelleny and Sacko (2001) introduced several new variables that were scaled together to create the CIRI Empowerment Index.[9] The CIRI empowerment variables are also listed and described in Table 1 and are discussed at length in Richards, Gelleny and Sacko (2001). Each index is based on the human rights reports from the US State Department and Amnesty International for all countries each year.

We use all observations from 1981 to 2009, for a total of 4,518 country-years. Each CIRI human rights variable measures the level of violation on an ordinal scale where 2 indicates that the right is not violated, 1 indicates that the right is violated occasionally and 0 indicates that the right is violated frequently. If the reports provide information about the number of individuals affected by a given rights violation, then the following cut-offs are used:

Level 0: 50 or more occurrences
Level 1: 1 to 49 occurrences
Level 2: 0 occurrences

According to the coder guidelines, if an estimate of the number of violations is not be available then the following guidelines from the CIRI codebook (Cingranelli and Richards, 2012a) are used:

- Instances where violations are described by adjectives such as ''gross'', ''widespread'', ''systematic'', ''epidemic'', ''extensive'', ''wholesale'', ''routine'', regularly or likewise, are to be coded as a ZERO (have occurred frequently).
- In instances where violations are described by adjectives such as ''numerous'', ''many'', ''various'' or likewise, you will have to use your best judgment from reading through the report to decide whether to assign that country a ONE (have occurred occasionally) or a ZERO (have occurred frequently). Look for language indicating a pattern of abuses; often, these cases merit a ZERO.

Thus there is a certain level of precision implied by the CIRI coding guidelines, as argued by Wood and Gibney (2010). However, our model directly confronts this issue by estimating the uncertainty of the latent variable estimates of human rights, which we describe in detail below.

[8] The descriptions of the four individual physical integrity variables in Table 1 and the Physical Integrity Scale are taken from the CIRI codebook (Cingranelli and Richards 2012a) and discussed at length in Cingranelli and Richards 1999, 2010; Richards, Gelleny and Sacko 2001.

[9] We include data used by CIRI to construct the ''new empowerment'' scale. For more details on the original scale, see the codebook in Cingranelli and Richards 2012a.

TABLE 1 *Variables in Human Rights Scales*

Item	Explanation
Physical Integrity Items	
Disappearances	Lack of deliberate disappearances of citizens by the government
Extrajudicial Killing	Lack of political and other extrajudicial killings or unlawful deprivation of life
Political Imprisonment	Lack of imprisonment because of religious, political or other beliefs in a given year
Torture	Lack of torture and other cruel, inhumane or degrading treatment or punishment
Empowerment Items	
Association	Freedom to assemble freely and join interest groups or other organizations
Foreign Movement	Freedom of individuals to leave and return to their country
Domestic movement	Freedom of individuals to travel within their country
Speech	Freedom of speech and press
Electoral	Existence of free and fair elections
Religion	Freedom from restrictions on religious practices
Worker	Rights of workers to collective bargaining, prohibition on compulsory labor rights, acceptable hours and working conditions

Note: The sources for all variables are Cingranelli and Richards (1999, 2010, 2012a,b); Richards, Gelleny and Sacko (2001).

Model Parameterization

We assume that the observed indicators for each country-year are functions of a unidimensional latent variable that represents the level of respect for human rights. To make this assumption more realistic, we run separate models using the physical integrity and empowerment indicators. For each country-year observation, let i index the country and t index the year. For each model, there are J indicators $J = 1,\ldots, J$, each of which is ordinal with levels 0 (no respect), 1 (moderate respect) and 2 (full respect). Our goal is to estimate each θ_{it}, which is the latent level of respect for physical integrity or empowerment rights of country i in year t.

Let $i = 1, \ldots, N$ index cross-sectional units and $t = 1, \ldots, T$ index time periods. In each time period, we observe values y_{ij} for each of $j = 1,\ldots,J$ indicators for each unit. Each indicator is ordinal in nature and can take on K_j values. The responses to each of the items depend on a single latent variable θ_{it} which may vary across units and over time.

For each item, there is an "item discrimination" parameter β_j and a set of $K_j - 1$ difficulty cut-points $\left(\alpha_{jk}\right)_{k=1}^{k_j}$. For each item, $y_{itj} = k$ if $\alpha_{jk-1}\ \theta_{it}\beta_j + \varepsilon_{itj} < \alpha_{j,k}$, where ε_{itj} is an error term and $\alpha_{j0} = -\infty$ and $\alpha_{jk_j} = \infty$ for notational convenience. The theoretical interpretation of the error term depends on the application. In the case of survey responses, individuals' responses to the questions may depend stochastically on the true latent variable. The applications we present below are based on human coding of documents, so it is natural to assume that ε_{itj} represents perceptual error on the part of the coders and reporting error on the part of the organizations that collect and aggregate the information that make up the documents. Obtaining a likelihood for this model depends on specifying a distribution for the ε_{itj} terms. In this case, we assume that they are independently drawn from a logistic distribution.

It follows that the probability distribution for a given response to item j is given by:

$$P\left[y_{itj} = k\right] = F\left(\alpha_{jk} - \theta_{it}\beta_j\right) - F\left(\alpha_{jk-1} - \theta_{it}\beta_j\right),$$

where $F(\cdot)$ denotes the logistic cumulative distribution function. Therefore, assuming local independence of responses across units, the likelihood function[10] for β, α and θ, given the data is:

$$\mathscr{L}(\beta, \alpha, \theta | y) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=1}^{J} \left[F\left(\alpha_{jy_{itj}} - \theta_{it}\beta_j\right) - F\left(\alpha_{jy_{itj}-1} - \theta_{it}\beta_j\right) \right]. \quad (1)$$

If θ was fully observed, the likelihood function above would be equivalent to independent ordinal logistic regression models. Since θ is unobserved, we are faced with the task of estimating the latent data along with the item-discrimination parameters and difficulty cut-points.

The products in Equation 1 reflect the local independence assumption utilized in some form by all IRT models. This implies that any two item responses are independent conditional on θ. That is, two item responses are only related because they measure the same latent variable. In this model, there are three relevant local independence assumptions, each of which is related to one of the products taken in Equation 1. The assumptions are: (1) local independence of different indicators within the same country-year, (2) local independence of indicators across countries within years and (3) local independence of indicators across years within countries.

Assumption 1 is strong in that it rules out causal relationships between items. For example, we assume that torture cannot cause an increase or decrease in political imprisonment, so any correlation between these two items is explained by the latent respect for physical integrity rights within a country. Though this assumption is made explicit in IRT models, we are unaware of existing human rights research that does not implicitly make an equally strong independence assumption. For instance, if a researcher runs a separate regression on each indicator to reach a broad conclusion about human rights practices, independence of errors is required to obtain efficient estimates. Furthermore, the additive index approach, based on Cingranelli and Richards' Mokken Scaling Analysis, is an IRT-based model that makes the identical conditional independence assumption that the (static) ordinal IRT model makes.[11]

[10] Note that the observed values of the indicators are incorporated into the definition of the likelihood function via the subscript y_{itj} and the subscript $y_{i,t,j-1}$ on the α terms.

[11] This assumption could be unreasonable; we are open to that possibility and are interested in pushing the model further to assess whether this is the case. We are currently aware of two papers that disaggregate the CIRI physical integrity scale and analyze some or all components jointly (as opposed to, say, projects that study just torture). One such study is Fariss and Schnakenberg (2013), which looks at systemwide co-occurances between different CIRI rights. A working paper by Conrad and Demeritt (2011) focuses on extrajudical killing and political imprisonment. These authors make this choice because "disappearances are ambiguous by their very nature" and "government torture can be used in conjunction with both state-sponsored killing and political imprisonment and strikes us as a complementary violation rather than one offering the possibility of substitution" (Conrad and Demeritt 2011, 14). The evidence presented by Conrad and Demeritt (2011) that state leaders choose to substitute one type of abuse for another is consistent with the assumption of our model: that the relationship is not directly causal but is instead dependent on the underlying latent trait, which is directly affected by the strategy of the leader. We are sure that there are additional working papers on this subject of which we are not yet aware. We simply wish to point out that these are important theoretical and empirical questions that the human rights literature is currently grappling with, and we believe that the model developed in our article can be extended to help address some of these important issues.

Assumption 2 may be problematic in many time-series cross-sectional data. Time period-specific shocks may cause increases or decreases in responses across many countries in the same year. In human rights data, however, the influence of these events on our estimates will be small, since within-country variation in our indicators is small compared to the between-country variation.

We depart from standard measurement models in the literature by relaxing Assumption 3. We argue that Assumption 3 is most problematic in many time-series cross-sectional datasets. For instance, if the behavior caused by the latent variable tends to persist over time, the local independence assumption is violated and researchers should expect biased estimates of the latent variable. We account for such dynamics while maintaining the basic form for Equation 1 by incorporating temporal information into prior beliefs about θ, as we describe below.

For the purposes of comparison, we estimate our dynamic model alongside the standard static IRT model described in Treier and Jackman (2008). The difference between the standard ordinal IRT (O-IRT) model and the dynamic ordinal IRT (DO-IRT) model lies in the specification of the hierarchical priors for θ_{it}. For the O-IRT model, we place independent standard normal priors on each θ_{it}. In other words,

$$\theta_{it} \sim N(0, 1)$$

for all i and t. For the DO-IRT models, we use the same standard normal prior when $t = 1$ and

$$\theta_{it} \sim N(\theta_{it-1}, \sigma)$$

for all other years. This method of incorporating dynamics was implemented in the context of a dichotomous IRT by Martin and Quinn (2002). One difference between our model and the Martin and Quinn model, besides our ordinal implementation, is that we estimate σ instead of specifying it *a priori*.

The prior for variance σ is modeled as $U(0,1)$. This reflects our prior knowledge that the between-country variation in human rights respect will be much higher on average than the average within-country variance.[12] Slightly informative gamma priors $Gamma(4,3)$ were specified for the β parameters. The prior on β has strictly positive support to reflect our prior belief, based on the Cingranelli and Richards (1999) article, that all indicators contribute significantly (and in the same direction) to the latent variable.[13] The α parameters were given $N(0,4)$ priors (extremely diffuse for this model) subject to the ordering constraint that $\alpha_{j1} > \alpha_{j2}$ for all j. Table 2 summarizes the parameterization of the DO-IRT and O-IRT models. As should be clear, the key difference between the two models is the prior distribution of the latent variable θ_{it}.

Note that, as is generally true of item-response models, the likelihood function in Equation 1 is not identified. In particular, IRT models suffer from "invariance to reflection", which means that (for instance) multiplying all of the parameters by -1 would have no effect on the likelihood function. Though lack of identification is problematic in maximum likelihood models, it is, in principle, not a problem for Bayesian approaches.

[12] Sensitivity checks reveal that this was not a consequential decision. Furthermore, the estimates of σ from the posterior of the converged model illustrate that the distribution is nowhere near 1, so the truncation decision was not important.

[13] Results from prior sensitivity analyses suggested that this was not extremely restrictive: when diffuse normal priors were specified for each β, the posterior densities rarely overlapped with 0.

TABLE 2 *Summary of DO-IRT and O-IRT Model Paramaterization*

	DO – IRT	O – IRT
Prior Distributions of Model Parameters		
Country-year latent variable (all years)	– – –	$\theta_{it} \sim N(0,1)$
Country-year latent variable (first year)	$\theta_{i1} \sim N(0,1)$	–- –- –-
Country-year latent variable (other years)	$\theta_{it} \sim N(\theta_{it-1},\sigma)$	–- –- –-
Uncertainty of latent variable	$\sigma \sim U(0,1)$	$\sigma \sim U(0,1)$
Item difficulty parameter	$\alpha_{jk} \sim N(0,4)$	$\alpha_{jk} \sim N(0,4)$
Item discrimination parameter	$\beta_j \sim$ *Gamma* (4,3)	$\beta_j \sim$ *Gamma* (4,3)
Probability Distribution $P[y_{it\ j} = k] =$ $F(\alpha_{jk} - \theta_{it}\beta_j) - F(\alpha_{jk-1} - \theta_{it}\beta_j)$	equivalent	equivalent
Likelihood $\mathscr{L}(\beta, \alpha, \theta \mid y) =$ $\prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{j=1}^{J}\left[F\left(\alpha_{jy_{itj}} - \theta_{it}\beta_j\right) - F\left(\alpha_{jy_{itj}-1} - \theta_{it}\beta_j\right)\right]$	equivalent	equivalent

The problem of invariance to rotation did, however, lead to some computational difficulties when estimating the model, which were eliminated by giving the β parameters strictly positive priors. For more information on identification problems in IRT models, see Jackman (2009).

Each model is estimated using three MCMC chains. Each chain is run with 300,000 iterations. The first 50,000 iterations were thrown away as burn-in and the rest were used for inferences. The Gibbs sampler for the DO-IRT and O-IRT models was implemented in Martyn Plummer's JAGS software (Plummer 2010). The JAGS code used is displayed in the appendix. The conventional diagnostics all suggested convergence,[14] including those of Geweke (1992), Heidelberger and Welch (1981, 1983), Gelman and Rubin (1992) and standard graphical analysis.

RESULTS

The model produces two sets of parameter estimates. The first set is item specific, and helps us make inferences about the relative informativeness of each indicator. The second set is the latent variable estimates for each observation, which are of primary interest. We review both sets of parameter estimates here. Though our DO-IRT and O-IRT models are estimated for both datasets, and item-specific parameters are displayed for both models, our discussion will focus on the better-fitting DO-IRT model; direct comparisons are saved for the following section of the article.

Results for Individual Indicators

The item-discrimination parameters (β) for each model allow assessment of the information value of each indicator. Figure 1 displays the item-discrimination parameters from the DO-IRT model for both physical integrity rights and empowerment rights. As expected, all items discriminated well with respect to the latent variable. The most informative indicator among physical integrity rights is extrajudicial killing (posterior mean: 3.541, 95 percent

[14] See Gill (2007) for a review of issues related to convergence diagnostics.

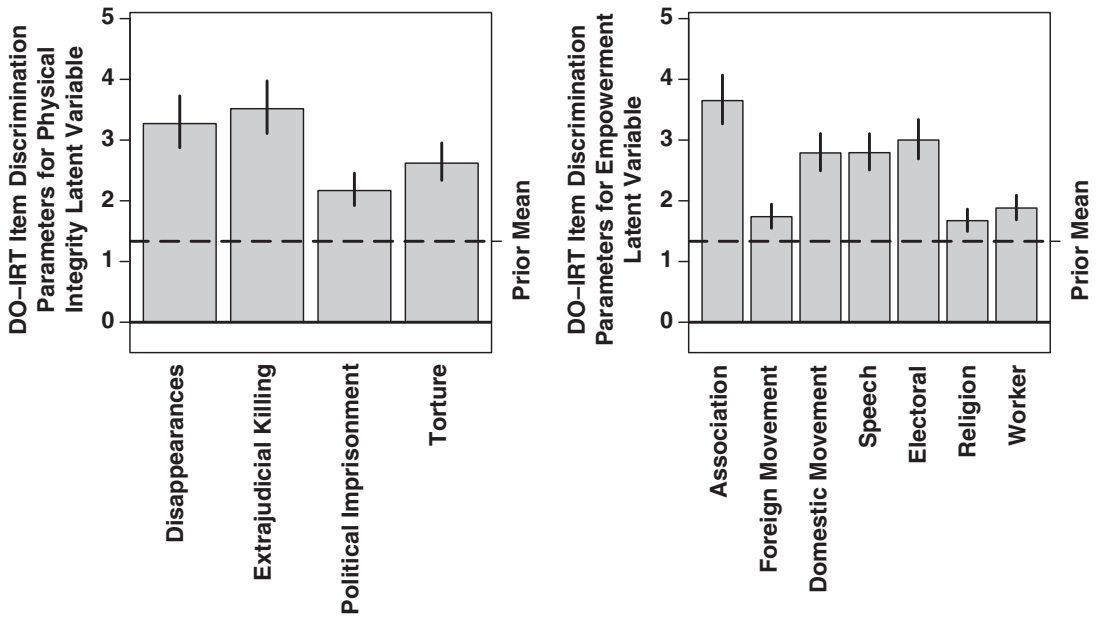

Fig. 1. Posterior densities for item-discrimination parameters for individual rights across models (30,000 draws)
Note: All of the item-discrimination parameters are different from the prior mean of the parameter.

credible interval [3.134, 3.976]), followed by disappearances (3.296: [2.889, 3.722]), torture (2.641: [2.356, 2.940]) and political imprisonment (2.186: [1.934, 2.444]).

Among the empowerment indicators, freedom of association (posterior mean: 3.639, 95 percent credible interval [3.259, 4.068]), followed by electoral self-determination (2.993: [2.690, 3.330]), freedom of speech (2.785: [2.499, 3.096]), domestic movement (2.78: [2.489, 3.104]), workers' rights (1.876: [1.688, 2.086]), foreign movement (1.734: [1.548, 1.937]) and freedom of religion (1.669: [1.496, 1.861]). The wide range of item-discrimination parameters for both models provides substantial evidence that our model improves on the additive scale used throughout most of the human rights literature. The item-discrimination parameters (β) and cut-points (α) for the ordinal indicators from the IRT models are displayed in Tables 3 and 4.

Finally, recall that we departed from previous dynamic models by estimating σ, the variance of the autoregressive prior, rather than specifying it *a priori*. The σ parameter gives a rough idea of the relative importance of within-country versus between-country influences variance. That is, since the overall variance is fixed at 1, a variance of 0 would indicate that respect for human rights does not change at all within countries, while a variance of 1 would suggest that respect for human rights within countries is just as variable as between countries. Recall that σ was restricted to be less than or equal to 1 based on prior knowledge. In the Physical Integrity model, the posterior mean for σ was 0.039 with a 95 percent credible interval from 0.031 to 0.049. In the Empowerment model, the posterior mean for σ was 0.034 with a credible interval from 0.026 to 0.043. These suggest that between-country variation is much larger than within-country variation, which helps to explain why the DO-IRT model performed better in posterior predictive checks. These statistics also suggest that the requirement that $\sigma \leq 1$ was not very restrictive.

TABLE 3 *Point Estimates and 95 percent Posterior Intervals for Item Discrimination Parameters β and Ordinal Cut-points α for Physical Integrity Rights Indicators from the DO-IRT and O-IRT Models*

Physical Integrity Items		DO – IRT		O – IRT	
Disappearances					
	β	3.296	[2.889, 3.722]	2.951	[2.688, 3.236]
	α_1	−5.379	[−5.945, −4.837]	−4.641	[−4.987, −4.320]
	α_2	−3.206	[−3.719, −2.708]	−2.443	[−2.667, −2.236]
Extrajudicial Killing					
	β	3.541	[3.134, 3.976]	3.444	[3.139, 3.785]
	α_1	−4.052	[−4.598, −3.520]	−3.450	[−3.757, −3.178]
	α_2	−0.893	[−1.393, −0.393]	−0.128	[−0.274, 0.014]
Political Imprisonment					
	β	2.186	[1.934, 2.444]	1.717	[1.596, 1.843]
	α_1	−1.826	[−2.142, −1.505]	−1.244	[−1.347, −1.144]
	α_2	0.056	[−0.250, 0.368]	0.448	[0.356, 0.538]
Torture					
	β	2.641	[2.356, 2.940]	2.426	[2.248, 2.616]
	α_1	−1.321	[−1.693, −0.942]	−0.745	[−0.862, −0.631]
	α_2	2.187	[1.815, 2.573]	2.535	[2.369, 2.710]
DIC		23,779		29,287	

TABLE 4 *Point Estimates and 95 percent Posterior Intervals for Item Discrimination Parameters β and Ordinal Cut-points α for Empowerment Rights Indicators from the DO-IRT and O-IRT Models*

Empowerment Items		DO – IRT		O – IRT	
Association					
	β	3.639	[3.259, 4.068]	3.734	[3.488, 3.999]
	α_1	−2.106	[−2.600, −1.606]	−2.052	[−2.239, −1.873]
	α_2	1.106	[0.627, 1.613]	0.984	[0.830, 1.142]
Foreign Movement					
	β	1.734	[1.548, 1.937]	1.689	[1.578, 1.803]
	α_1	−2.590	[−2.847, −2.333]	−2.516	[−2.644, −2.389]
	α_2	−0.810	[−1.048, −0.566]	−0.799	[−0.895, −0.705]
Domestic Movement					
	β	2.78	[2.489, 3.104]	2.605	[2.448, 2.771]
	α_1	−3.317	[−3.713, −2.912]	−3.078	[−3.253, −2.909]
	α_2	−0.361	[−0.729, 0.02]	−0.355	[−0.472, −0.239]
Speech					
	β	2.785	[2.499, 3.096]	3.070	[2.882, 3.27]
	α_1	−2.134	[−2.519, −1.745]	−2.224	[−2.391, −2.063]
	α_2	2.096	[1.72, 2.488]	2.047	[1.889, 2.211]
Electoral					
	β	2.993	[2.690, 3.330]	3.350	[3.139, 3.568]
	α_1	−1.943	[−2.350, −1.529]	−2.041	[−2.208, −1.879]
	α_2	0.809	[0.411, 1.227]	0.769	[0.629, 0.908]
Religon					
	β	1.669	[1.496, 1.861]	1.706	[1.602, 1.816]
	α_1	−1.846	[−2.081, −1.607]	−1.829	[−1.938, −1.724]
	α_2	−0.051	[−0.274, 0.182]	−0.086	[−0.173, 0.001]
Worker					
	β	1.876	[1.688, 2.086]	1.875	[1.767, 1.987]
	α_1	−1.116	[−1.37, −0.853]	−1.105	[−1.204, −1.006]
	α_2	1.683	[1.425, 1.953]	1.551	[1.447, 1.658]
DIC		41,840		45,651	

Estimates of Latent Human Rights Levels

The primary results from our statistical analysis are the estimates of latent respect for physical integrity and empowerment rights for each country-year. Since the 4,518 estimates provided at each draw in the model are difficult to display in their entirety, we present illustrative slices of the data and note that the entire set of estimates, with associated posterior standard deviations, is available online.

Figures A1 and A2 in the appendix display the latent variable estimates for physical integrity and empowerment rights, respectively, with 95 percent credible intervals, for all countries in the data in 2008. The range of parameter estimates for both models is around -3 to 3, and the credible intervals cover about one unit for most of the observations. As with confidence intervals, one should not judge the statistical significance of differences between countries by examining the overlap of credible intervals (Schenker and Gentleman 2001), but a more systematic assessment of differences between countries is given later.

Our model also enhances researchers' ability to assess change in human rights levels over time. Figure 2 displays two columns of plots for three different countries (China, Guatemala and Namibia). The plots in the left column display the highest and lowest country-year posterior densities for each of these countries, and the right column displays the mean estimates and 95 percent confidence intervals for every year for these countries. The plots on the rights column provide a qualitative assessment of the human rights trends in each of the three countries. Visual examination of the plots reveals a slow decline in human rights levels in China, despite starting from a very low level. In contrast, the plots show radical improvement in human rights levels in Namibia and substantial improvement in Guatemala over the time period.

The plots on the left column of Figure 2 help us make more systematic comparisons between particular years within countries. We chose the highest and lowest years for analysis. In addition, a statistical comparison of the draws from the model allows us to give the equivalent of a p-value for the hypothesis of a difference between the highest and lowest years. The plot shows unambiguous differences between high and low years for Guatemala ($p < 0.001$) and Namibia ($p < 0.001$) and substantial differences between high and low years for China ($p < 0.01$).

Comparison with Traditional Measures

Since our model estimates produce latent variable estimates that researchers can use to replace traditional additive scales, we provide a few revealing comparisons between our model and the traditional measures. Though the two measures are highly correlated, as should be expected, our model-based estimates provide a more nuanced picture of global human rights practices in several ways.

We address the issue of discrimination between countries. An advantage of our latent variable estimates is that they allow the researcher to assess the level of error in the measure. This has been an issue of concern to quantitative human rights researchers. Wood and Gibney (2010) argue that the method of aggregating multiple types of abuse into one scale implies a level of precision that is not supported by the data. Our method is a response to this objection. Furthermore, systematic comparison of our measure to the traditional measures confirms the validity of this objection to the current scales.

It is important to emphasize that there is no model-free way to estimate latent levels of respect for human rights. Thus the additive scale approach is a *model* that assumes equally weighted indicators and no error. If two country-years have the same value on the CIRI

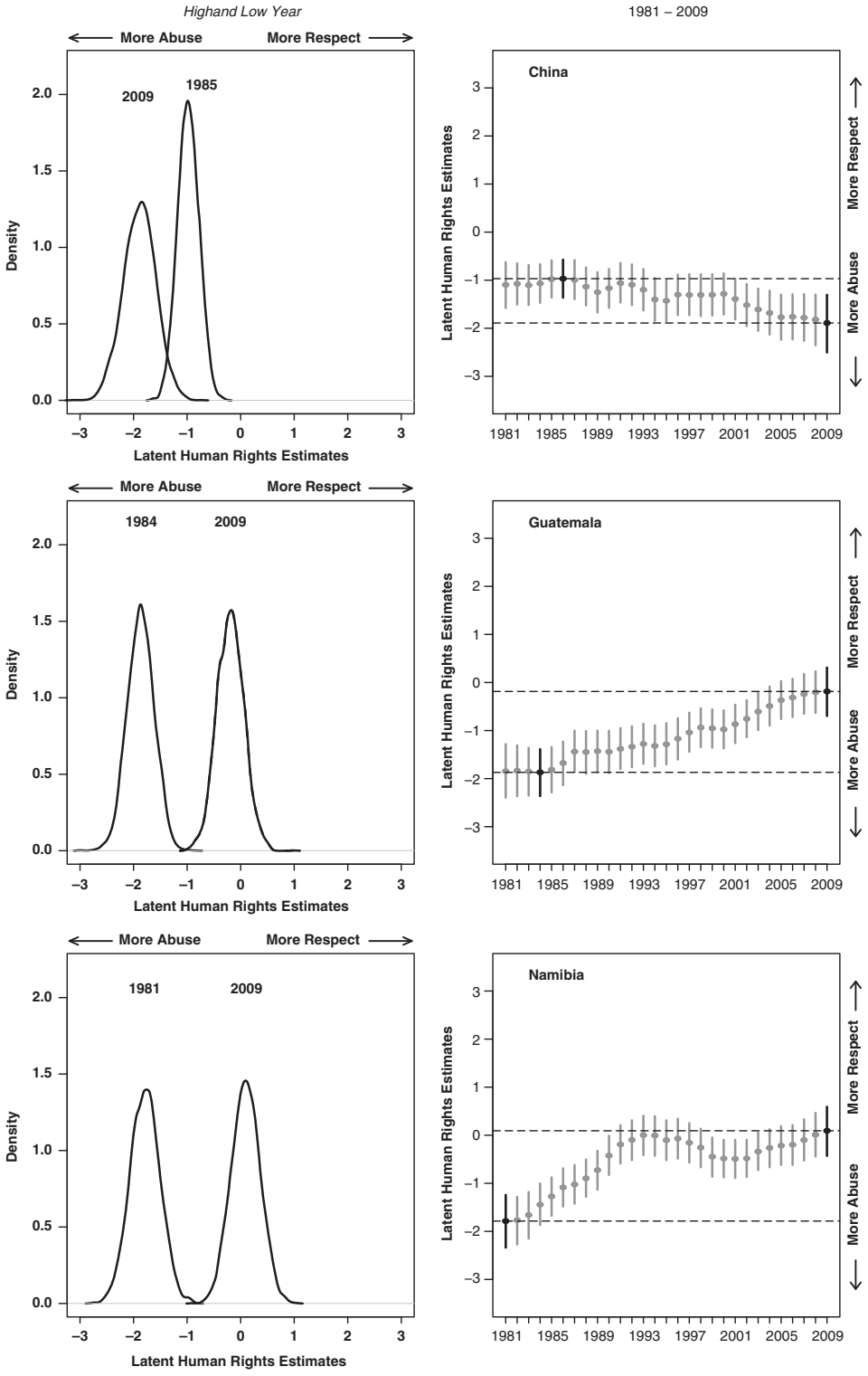

Fig. 2. Posterior estimates from three country examples
Note: DO-IRT highest and lowest posterior density for China, Guatemala and Namibia are displayed in the left column of plots. In the right column of plots, the dots are point estimates (posterior means) and the lines are 95 percent credible intervals from 1981-2009.

TABLE 5 *Precision from DO-IRT Model versus the CIRI Physical Integrity Index*

CIRI Value	0.9	0.95	0.99	N
0	0.330	0.240	0.122	230
1	0.257	0.177	0.078	250
2	0.210	0.146	0.067	332
3	0.254	0.173	0.071	402
4	0.206	0.135	0.055	641
5	0.206	0.135	0.055	641
6	0.261	0.181	0.079	640
7	0.262	0.183	0.085	759
8	0.411	0.323	0.185	684

Note: The column values are the proportion of the country-year pairs for which both country-years in the dyad received the same score on the CIRI physical integrity index but for which one is greater than the other on the DO-IRT latent variable with 0.90, 0.95 and 0.99 probability, respectively.

TABLE 6 *Precision from DO-IRT Model Versus the CIRI Empowerment Index*

CIRI Value	0.90	0.95	0.99	N
0	0.300	0.215	0.103	149
1	0.330	0.240	0.123	170
2	0.284	0.194	0.091	182
3	0.206	0.132	0.048	252
4	0.199	0.131	0.052	273
5	0.112	0.062	0.019	317
6	0.153	0.085	0.024	336
7	0.159	0.100	0.039	290
8	0.179	0.113	0.040	226
9	0.227	0.151	0.058	244
10	0.191	0.122	0.045	296
11	0.218	0.145	0.059	336
12	0.206	0.132	0.051	423
13	0.269	0.193	0.095	516
14	0.390	0.306	0.179	508

Note: The column values are the proportion of the country-year pairs for which both country-years in the dyad received the same score on the CIRI empowerment index but for which one is greater than the other on the DO-IRT latent variable with 0.90, 0.95 and 0.99 probability, respectively.

additive scale, the additive scale model states that those country-years are the same with a probability of 1. Our model finds substantial evidence of variation in latent respect for human rights within levels of the traditional additive scale. Table 5 displays, for each level of the physical integrity additive scale, the proportion of country-year pairs with that value, such that one country-year is greater than the other with high probability. Table 6 displays the same information for the Empowerment Index.

The DO-IRT model applied to the physical integrity variables produces significant variation within values of the CIRI additive scale. For every level of the additive scale, over 20 percent of the pairs were different with a probability greater than 0.9. The higher number of differences on the 0 and 8 ends of the scale reflect a conservativeness built into the dynamic model: a country that suddenly experiences more extreme values in a given year will have a higher variance and a more moderate posterior than one that had extreme values for the entire time period.

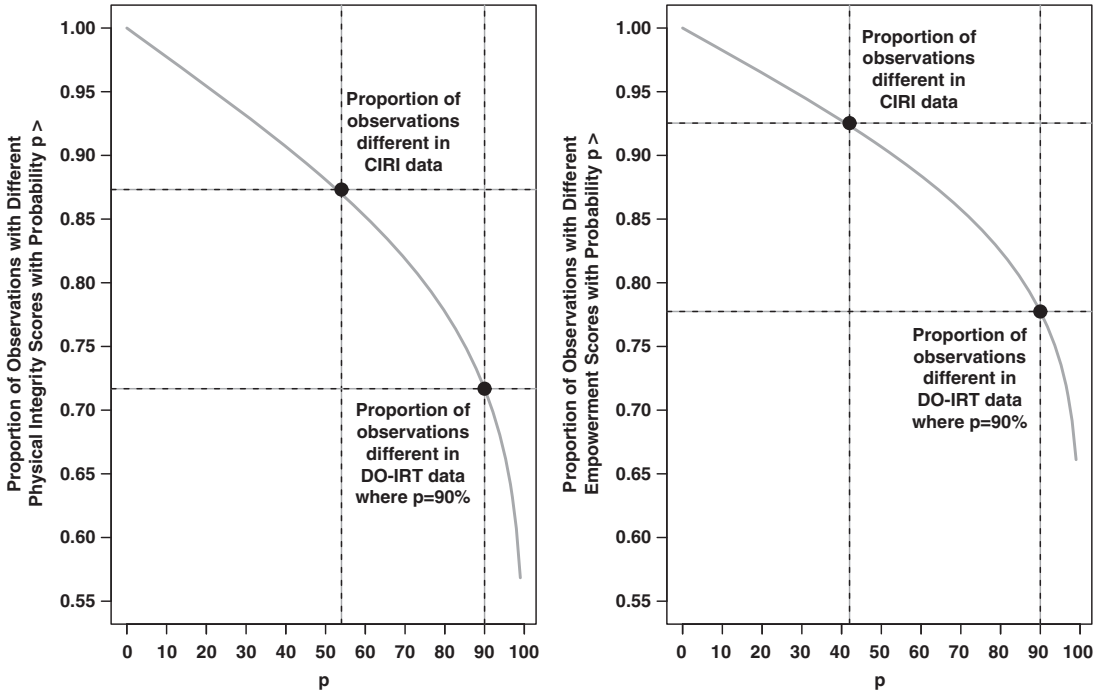

Fig. 3. Summary of paired country-year comparisons for all such pairs in the DO-IRT physical integrity model and empowerment model
Note: The Y-axis in each graph represents the proportion of country-year pairs that the DO-IRT model predicts are different from one another with level of confidence p on the x-axis.

Similar patterns can be seen in the empowerment indicators. For all values of the Empowerment Index, over 11 percent of the country-year pairs are different with probability greater than 0.9, and the number is greater than 20 percent for the majority of index levels. As in the physical integrity data, larger differences are seen at extreme values of the scale.

Though the DO-IRT model reveals substantial variation between country-years that are considered the same under the additive scale model, it is more modest than the additive scale model in terms of its overall ability to discriminate between country-years. Figure 3 displays an overall picture of country-year comparisons for the physical integrity and empowerment data. Notice that the CIRI data assumes that approximately 90 percent of these country-year pairs are different with probability 1 and the remaining are different with probability 0. Our model suggests that this proportion of country-year pair differences is not very likely.

These results suggest that in many cases, scholars will get a very different picture of comparisons of human rights behavior at the individual-case level from looking at our model. Each of the country-year comparisons represented in our tables represents different descriptive inferences that one might wish to draw. Is Uzbekistan more repressive today than when it first emerged as an internationally recognized state? In 2008, was Tajikistan more repressive than the Kyrgyz Republic? Did repression decrease in Guatemala immediately after the Cold War? We can now estimate that the probability of a "yes" answer to these questions is 0.586, 0.743 and 0.639, respectively.[15]

[15] Though the first two comparisons are simple country-year comparisons, the last comparison is computed by taking the average value for Guatemala 1988–91 and comparing it to the average value in 1992–95 for each

MODEL CHECKS

The quality of the descriptive inferences we make about human rights levels in different countries depends critically on choosing the best model specification available. The DO-IRT model provides a better fit in a variety of ways, and provides a more realistic picture of changes in human rights practices. We compare the O-IRT and DO-IRT models in two ways. First, we use a formal decision-theoretic criterion called the Deviance Information Criterion to compare the models. Secondly, we provide a variety of posterior predictive checks, particularly related to dynamic aspects of human rights patterns.

Deviance Information Criterion

The O-IRT and DO-IRT models both fit the data reasonably well and are based on defensible assumptions about the data. To choose between models, more formal model comparisons are needed. Bayes Factors are often viewed as a good way to compare models in the Bayesian framework (Kass and Raftery 1995). Calculation of Bayes Factors, which requires marginalization over the parameter space of the models, remains difficult for very high dimensional models such as ours. A recent measure of model fit by Spiegelhalter *et al.* (2002)—called the Deviance Information Criterion (DIC)—is appropriate for comparing IRT models, is much easier to compute and is designed explicitly for models estimated using MCMC.

The DIC is an estimate of expected deviance and has been proposed as a measure of model fit when the goal is to maximize out-of-sample predictive power (Gelman *et al.* 2003). For a given factor of parameters ψ, the deviance is given by $D(y, \psi) = -2\log(\mathscr{L}(y|\psi))$, where $\mathscr{L}(y|\Psi)$ is the likelihood function of the model. Other commonly used information criteria use the number of parameters as an argument, but in a hierarchical context (such as in our DO-IRT model) the number of parameters can be difficult to quantify. The DIC uses the *effective number of parameters* which is:

$$pD = \overline{D}(y) - \hat{D}\left(y, \hat{\Psi}\right),$$

where $\overline{D}(y)$ is the posterior mean of the deviance and $\hat{D}\left(y, \hat{\Psi}\right)$ is the deviance estimates using the posterior mean of the parameters, $\hat{\Psi}$. The DIC is:

$$DIC = 2\overline{D}(y) - \hat{D}\left(y, \hat{\Psi}\right).$$

The model with the smallest DIC is expected to have the greatest out-of-sample predictive power. In addition to computational ease, the DIC carries some advantages over computing Bayes Factors. Bayes Factors may only be appropriate under the assumption that one and only one of the models is ''true'' and the goal is to choose the true model (Spiegelhalter *et al.* 2002). This assumption is especially unappealing for IRT models, since the latent variable is a figment of our imaginations, used for the specific purpose of reducing the data into a lower dimensional space. We believe that criteria based on short-term predictive power, such as the DIC, are therefore more appropriate

(F'note continued)
draw and recording the number of times that the first average is lower than the second. Comparing the same Cold War time period to a later time period (1996–99) reveals a higher probability of difference (0.835).

for choosing between item-response models. A model that would best predict a new set of observations is likely to be most informative about the latent variable of interest. Like Bayes Factors and the other standard information criteria, the DIC penalizes more complex models, so parsimonious models are favored, all else equal.

The DIC for our models estimated on the human rights data suggests that DO-IRT performs better that O-IRT for both the physical integrity and empowerment data. The DIC for the DO-IRT model estimated usin the physical integrity data was 23,779 compared to 29,287 for O-IRT. Similarly, the DIC for the DO-IRT model applied to the empowerment data was 41,840 compared to 45,651 for O-IRT. There is no objective standard for what constitutes a substantial difference in the DIC, but suggested ''rules of thumb'' (Spiegelhalter *et al*. 2002) propose that differences of greater than five or ten provide substantial evidence in favor of the model with the lower DIC. Though these rules of thumb are arbitrary, we point out that in our model comparisons, with differences of a few thousand in both cases, the rules of thumb are far from binding. Thus the DIC provides good evidence in favor of the DO-IRT model. Along with the concerns about the local independence assumptions discussed in Section 3, this leads us to recommend the DO-IRT model for use by human rights researchers.

Posterior Predictive Checks

Posterior predictive checks provide an additional method of assessment of model quality (Gelman and Hill 2007) and provide more insight into the specific reasons for differences in fit. We test the ability of the DO-IRT and O-IRT models to predict the original CIRI response variables. For each draw from the posterior distribution, we predict each of the j items y_{itj} for every country-year observation it. We then calculate the sum of squared differences of observed y_{itj} and d posterior predicted values $\hat{y}_{itjd}$ using the following equation:

$$S_{itj} = \sum_{d} \left(y_{itj} - \hat{y}_{itjd}\right)^2.$$

For presentation purposes, we aggregate the sum of squares difference for each observation. Figure 4 displays a proportion for each item in which the average sum of squares differences for each item is calculated from the DO-IRT and O-IRT models.[16] The proportions increase as the number of observations that have a smaller sum of squared deviation increases when comparing the DO-IRT model to the O-IRT model. On average, there is slightly less deviation from the ''true'' CIRI item and the predicted item generated by the DO-IRT model compared to the predictions generated by the O-IRT model. The proportion columns in Table 7 measure the proportion of country-year observations that have a smaller deviation generated by the DO-IRT model compared to the O-IRT model. In sum, the DO-IRT model does a slightly better job of predicting the original CIRI data compared to the O-IRT model. However, we demonstrate next that the dynamic model (DO-IRT) does substantially better at predicting temporal changes compared to the static version (O-IRT).

We also assess the ability of the DO-IRT and O-IRT models to predict the temporal dynamics of the CIRI data. To accomplish this task, we repeated the procedure outlined

[16] These are highly accurate estimates: 9,000 posterior draws were used to generate these statistics.

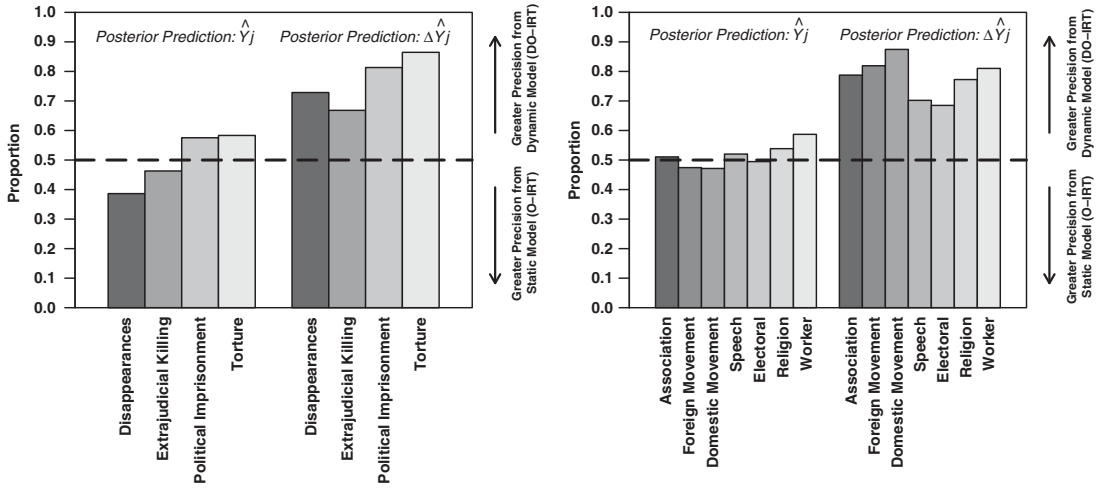

Fig. 4. Proportion of observations for which the Physical Integrity (left)/Empowerment (right) DO-IRT is more precise than the O-IRT estimate

Note: proportions closer to 1 indicate that the dynamic version of the model is outperforming the static version at predicting the original CIRI items and changes in those items from year $t-1$ to year t. Proportions closer to 0 indicate that the static version of the model is outperforming the dynamic version. Proportions at 0.50 indicate that both models are predicting the items with about the same amount of error. Notice that while a few of the proportions in the first figure are below the 0.50 mark, only in the case of predicting disappearances does the static model (O-IRT) substantially outperform the dynamic model. The dynamic model is clearly superior at predicting temporal changes in the original data.

TABLE 7 *Summary of Posterior Predictive Checks*

Items	Predicted Item	Predicted Δ Item
Physical Integrity Items		
Disappearances	0.386	0.729
Extrajudicial Killing	0.463	0.668
Political Imprisonment	0.574	0.813
Torture	0.583	0.865
Average	**0.502**	**0.769**
Empowerment Items		
Association	0.511	0.788
Foreign Movement	0.474	0.819
Domestic movement	0.471	0.875
Speech	0.520	0.702
Electoral	0.495	0.685
Religion	0.539	0.772
Worker	0.587	0.810
Average	**0.514**	**0.779**

Note: The first column of proportions measures the proportion of country-year observations that have a smaller sum of squared deviation generated by comparing the observed item and predicted item for the DO-IRT model compared to the O-IRT model. The second proportion measures changes from year $t-1$ to year t for country-year observations that have a smaller sum of squared deviation generated by the DO-IRT model compared to the O-IRT model. In sum, the DO-IRT model does a slightly better job of predicting the item and a substantially better job of predicting temporal change from year $t-1$ to year t compared to predictions generated from the O-IRT model.

above, using the first differences of the observed data and the first differences of the predicted data taken from the posterior draws. For each draw from the posterior distribution, we predict the change in each of the j items y_{itj} for every country-year observation it. We then calculate the sum of squared differences of observed Δy_{itj} and d posterior predicted values $\Delta \hat{y}_{it jd}$ using

$$\Delta S_{it j} = \sum_{d} \left(\Delta y_{it j} - \Delta \hat{y}_{it jd} \right)^2.$$

Overall, the results displayed in both Table 7 and Figure 4 demonstrate the improvement in predictive power of the DO-IRT model when compared to the O-IRT model. The DO-IRT model outpredicts the O-IRT model for both sets of predictions. However, it substantially outperforms the predictive power of the O-IRT model when considering changes in time.

Though the dynamic model is clearly superior at predicting temporal changes in the original data, this result should not be surprising, since the model explicitly includes temporal information in the prior of each estimate of the latent variable. We explore additional temporal predictions next, and discuss the predictive validity of the dynamic and static measures.

Predictive Validity

A measure of the construct validity of our measure is the performance of the human rights variable at predicting future human rights levels. The difficulty in assessing the relative predictive abilities of the variables is in selecting the outcome to predict. Since most measures should perform better at predicting future values of themselves than of another measure of the same concept, the choice of outcome variable may bias the comparison in favor of one measure or another. To make the comparison as difficult as possible for the DO-IRT model, we compared the ability of each variable to predict future values of the traditional CIRI additive index. To do this, we regressed the CIRI Physical Integrity Scale on lagged values of each variable and compared the model deviance and sum of squared deviations resulting from each model. Specifically, we regress these scales on (1) the lagged version of the scales, (2) the lagged version of the static latent variable and (3) the lagged version of the dynamic latent variable. We also repeat Models 2 and 3 in order to incorporate the uncertainty captured by the latent variables in the predictions generated by these models. We describe these procedures at the end of this section. The results of this comparison are presented in Table 8.

For each of the models using the latent variable estimates, we ran one regression using the posterior mean as a point estimate of the human rights index in each country-year and another taking 1,000 draws from the posterior to account for the uncertainty in the estimates. To incorporate the uncertainty from the measurement models, we estimate the same statistical model for each of the m datasets and then make inferences from the distribution of parameter estimates.

Table 8 displays the model deviance and sum of square deviations calculated for five ordered logistic regression models with a single covariate. Though the deviance is the more accepted method for assessing fit in generalized linear models, we will refer to the sum of squared deviations in the discussion for ease of interpretation, since both numbers give the same impression. The sum of squared deviation statistics is generated from

TABLE 8 *Ordered Logistic Regressions: CIRI Physical Integrity Additive Index Dependent Variable*

Model	Lagged Independent Variable	Sum of Squares [95% CI]	Model Deviance [95% CI]
1	DO-IRT Physical Integrity$_{t-1}$	3592	9707.4
2	DO-IRT Physical Integrity$_{t-1,d}$	4821 [4667.9, 4997.0]	10844.6 [10730.0, 10957.7]
3	CIRI Physical Integrity Index$_{t-1}$	6318	12165.1
4	O-IRT Physical Integrity$_{t-1}$	7247	12364.8
5	O-IRT Physical Integrity$_{t-1,d}$	10643 [10254.0, 11037.1]	13991.2 [13871.4, 14127.9]

Note: summary statistics are derived from five bivariate ordered logistic regressions in which the CIRI Physical Additive Integrity Index is regressed on one of several lagged physical integrity variables. The models are listed in ascending order of predictive fit. The mean estimate of the dynamic latent variable (Model 1) outperforms the alternative lagged constructs even when uncertainty is incorporated into the model (Model 2). Note that the lagged version of the original CIRI Physical Integrity Index (Model 3) outperforms the models with a static latent variable (Models 4 and 5).

TABLE 9 *Ordered Logistic Regressions: CIRI Empowerment Additive Index Dependent Variable*

Model	Lagged Independent Variable	Sum of Squares [95% CI]	Model Deviance [95% CI]
1	DO-IRT Empowerment$_{t-1}$	6007	12214.6
2	DO-IRT Empowerment$_{t-1,d}$	8508 [8197.0, 8848.0]	13653.23 [13517.5, 13776.5]
3	CIRI Empowerment Index$_{t-1}$	9377	14178.4
4	O-IRT Empowerment$_{t-1}$	10752	14659.9
5	O-IRT Empowerment$_{t-1,d}$	18689 [17864.0, 19514.2]	16922.17 [16769.8, 17068.6]

Note: summary statistics are derived from five bivariate ordered logistic regressions in which the CIRI Empowerment Additive Index is regressed on one of several lagged empowerment variables. The models are listed in ascending order of predictive fit. The mean estimate of the dynamic latent variable (Model 1) outperforms the alternative lagged constructs even when uncertainty is incorporated into the model (Model 2). Note that the lagged version of the original CIRI Physical Integrity Index (Model 3) outperforms the models with a static latent variable (Models 4 and 5).

ordered logistic regressions; the CIRI Physical Integrity Index is regressed on the lagged version of itself and several other lagged measures of physical integrity abuse. The lagged version of the CIRI Index does better than the mean estimate of the lagged static variable, but not the mean estimate of the lagged dynamic variable. As an alternative, we also ran the ordered logistic regressions 1,000 times, taking draws from the lagged mean and standard deviation of the dynamic and static estimates. Again the dynamic estimate 4,821 [95 percent CI: 4667.93,4997.00], even when incorporating uncertainty about the estimate from the previous year, outperforms both the models with the lagged CIRI value 6,318 and a model that also incorporates uncertainty from the static model 10,643 [95 percent CI: 10253.98, 11037.10]. We generated the same statics using the empowerment variables, which are displayed in Table 9. Again the dynamic latent variable outperforms both the lagged CIRI empowerment variable and the static estimate. Overall, the evidence provided in this section demonstrates that the dynamic latent variable model produces much more precise and informative estimates, which are useful in statistical analyses of human rights practices.

ILLUSTRATION: HUMAN RIGHTS AND TERRORISM

We now illustrate how to use our measure for applied data problems by repeating an analysis by Piazza and Walsh (2009), which shows that countries with better human rights practices experience fewer terrorist attacks. Piazza and Walsh use negative binomial models on counts of the number of domestic terrorist attacks, transnational terrorist attacks and total terrorist attacks that occurred in each country-year. The main independent variable of interest was the CIRI Physical Integrity Rights Index. As controls, the model also includes three measures of democracy: (1) constraints on the executive, participation and the durability of the regime; three measures of state capacity (government involvement in an international war, government involvement in a civil war and the natural logarithm of the state's population) and (3) the natural logarithm of gross domestic product (GDP) per capita. We replicated their analysis of domestic terrorist events and then repeated the analysis using our latent variable estimates in place of the additive scales.[17] Table 10 compares the original and revised models.

In the model using the latent variable estimates, we incorporated the uncertainty associated with the measure using the same Monte Carlo procedure that is used for multiply imputed missing data.[18] First, we take $m = 10$ draws from the posterior distribution of the DO-IRT model and use them to create 10 datasets. Next, estimates were combined in the same way they are combined in multiple imputation problems. Estimates were combined using the Rubin (1987) formulas, in which the point estimate for each parameter is the mean from the m estimates, and the standard error is:

$$\sqrt{\frac{1}{m}\sum_{k=1}^{m} s_k^2 + \left(1 + \frac{1}{m}\right)\sigma_\beta^2}$$

where s_k^2 is the standard error from dataset k, and σ_β^2 is the variance in the regression coefficients between datasets. In other words, the standard error is the average standard error from each model, plus the variance in the regression coefficients times a correction factor for finite m. This procedure can be implemented in any program designed for multiply imputed data.

Though the coefficients for the human rights variable in the original and revised models are different, these differences correspond primarily to the different scales for the variables. The substantive effects, and therefore the conclusions of Piazza and Walsh (2009), are very similar across the two models.[19] To compare substantive effects across the models, we simulated the expected number of attacks under both models. The change in the expected number of attacks when changing physical integrity from the 25th percentile to the 75th percentile was about −3.22 (95 percent CI: −4.71, −2.06) in the original model and −3.32 (95 percent CI: −4.60, −2.32) when including the latent variable. The main difference between these two effects is the size of the confidence interval of the

[17] We were able to precisely replicate the results in Piazza and Walsh (2009) using Stata. However, the replication in R produced slightly different coefficient estimates. Here we report information based on the replication in R, but the inferences we draw from this comparison are the same as those from the parallel analysis in Stata. We have chosen to discuss the results estimated in R because the procedure we demonstrate is more easily implemented in R than in Stata.

[18] Among others, Mislevy (1991) advocated this approach in the context of latent variable models.

[19] We should also note that our experimentation with the model in Piazza and Walsh (2009) revealed that their main findings are extremely robust to alternative specifications.

TABLE 10 *Negative Binomial Regression of the Number of Domestic Terrorist Events*

	Original Model		Revised Model	
Variable	Estimate	Standard Error	Estimate	Standard Error
(Intercept)	−7.706	(1.199)	−11.821	(1.312)
Physical Integrity	−0.406	(0.066)	−1.448	(0.210)
Participation	0.016	(0.008)	0.010	(0.009)
Executive Constraints	0.179	(0.088)	0.185	(0.089)
Durable	1.095	(0.259)	1.283	(0.265)
International War	0.710	(0.698)	0.544	(0.696)
Civil War	1.099	(0.385)	0.915	(0.425)
Population (ln)	0.606	(0.076)	0.542	(0.078)
GDP per capita (ln)	0.340	(0.122)	0.605	(0.130)
N	765		765	
ln $\mathscr{L}$	−1318.533		−1276.265	

estimated difference. A Kolmogorov-Smirnov test confirms that the two distributions of simulated estimates are statistically different ($p < 1e^{-6}$). Thus the increased precision of the DO-IRT estimates relative to the additive index allows us to estimate effects more precisely *even after we incorporate the uncertainty associated with measurement error*. The substantive impact of this observation is that researchers may be able to detect effects that would have been missed using additive scales.

The superior performance of the DO-IRT model can also be verified by comparing the fit of the models.[20] The log-likelihood in the original model is −1318.533. The average log-likelihood from the 10 replications is −1276.265. Note also that each of the individual log-likelihoods from the 10 replications are larger than the log-likelihood from the original model, which suggests that the model with the latent human rights variable better fits the data in each individual m model.

CONCLUSIONS

In this article, we introduced the Bayesian ordinal IRT model and provided a theoretical motivation for developing the dynamic version of this model when applied to human rights data. Overall, the Bayesian approach provides the researcher with a high degree of flexibility in modeling latent characteristics of states or any other political actor. Further, the approach utilized in this article has the potential to occupy an important middle ground in the current debate over measurement of respect for physical integrity rights.[21] Critics of the CIRI Physical Integrity Index are uncomfortable with several aspects of the measure. First, they are skeptical of the assumption that each practice represents the latent trait equally. Secondly, scholars are concerned that the level of precision implied by the estimates is not supported by the available data. Our approach addresses these concerns by empirically estimating both the item weights and the uncertainty of the estimates.

[20] Log-likelihood comparisons are sufficient for evaluating the relative fit of the models, since the models have the same dependent variable and the same number of parameters.

[21] See Cingranelli and Richards (2010), Wood and Gibney (2010) and more recently Clark and Sikkink (Forthcoming).

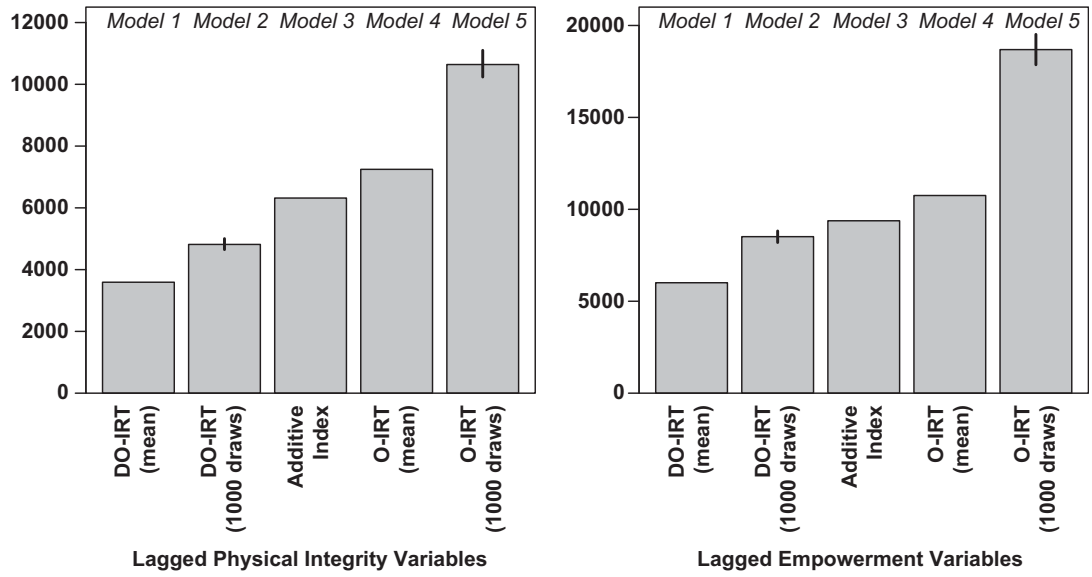

Fig. 5. Comparison of deviations from bivariate model predictions of CIRI Physical Integrity (left)/ Empowerment (right) additive index
Note: The sum of squared deviations is derived from each of five bivariate ordered logistic regressions in which the CIRI Physical Integrity Additive Index (left panel) and the CIRI Empowerment Additive Index (right panel) are regressed on one of five different lagged variables. Lower values represent a better-fitting model. The dynamic latent variable in period $t-1$ predicts values of the indices in period t with greater accuracy than the static variables or the additive indices themselves.

At the same time, our approach provides the advantage of utilizing a disaggregated data source that allows empirical assessment of unidimensionality and places the aggregation of indicators under the control of the researcher, thus allowing for greater transparency in both measurement and testing. Further, the point estimates and credible intervals estimated in this analysis are available for researchers to use as both a dependent variable or independent variable in applied work. These continuous latent human rights variables allow the researcher to use the widely available tools for assessing such a dependent variable in a panel data setting. We have also discussed two simple techniques for including the uncertainty inherent in measuring human rights in statistical analyses that include the latent human rights variables as an independent variable.

Clark and Sikkink (forthcoming) are critical of both the CIRI physical integrity data and the Political Terror Scale. These authors suggest that ''systematic 'information effects' in these data sets may contribute to the pessimistic findings of the quantitative literature'' (forthcoming). By pessimistic findings, these authors mean the stagnant pattern of human rights abuse over time and the negative correlation between these human rights scales and the implementation of UN human rights treaties. Our model does not allow us to assess information effects directly. However, the DO-IRT model is extendable and, as we have demonstrated in this article, capable of (1) bringing together diverse sources of information, (2) assessing the relative quality of the information included and (3) quantifying the certainty of estimates of repression that are generated from the models. Future measurement projects can use our DO-IRT model to

parameterize new theoretical insights. Additional data collection efforts are necessary to address the challenge posed by Clark and Sikkink (forthcoming), which is similar to one made by Stohl *et al*. (1986) three decades ago. Another issue that our model might address in future research is the comparability of event data that counts the number of repressive events in country-year observations.[22] We leave these extensions to future research projects.

Overall, the DO-IRT model provides a starting point for new theorizing and model development by generating new information about quantifiable country-year comparisons that was not previously available to researchers. The dynamic latent variable estimates generated by the DO-IRT model allow for the direct, probabilistic comparison of the level of human rights abuse between country-year cases. Until now, analysts who wished to compare two cases had to rely on case-specific evidence or ordinal human rights variables. Case-specific evidence limits the number of comparisons that can be made and does not provide an estimate that can be compared alongside other pairwise comparisons. The ordinal data that predominates the quantitative human rights literature can provide yes/no answers but is unable to quantify the certainty of a given answer. The evidence provided in Figures 4 and 5 and Tables 7, 8 and 9 all suggest that the dynamic latent physical integrity and empowerment variables provide the most informative and precise estimates of their corresponding theoretical constructs to date. This evidence, coupled with the ability to make probabilistic comparisons between country-year cases (which we demonstrated in Figure 3 and Tables 5 and 6) means that the estimates generated from this project can provide transparent and realistic answers to the descriptive questions we posed at the outset of this article. The estimates should be of use in both large-N and case-study research. However, we hope that it also proves useful to policy makers and NGOs. The ability to make probabilistic statements and accurate predictions about future levels of abuse is essential for scientific progress and targeted action by human rights NGOs.

REFERENCES

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

Aleman, Eduardo, and Sebastiáns M. Saiegh. 2007. 'Legislative Preferences, Political Parties, and Coalition Unity in Chile'. *Comparative Politics* 39(3):253–72.

Brysk, Alison. 1994. 'The Politics of Measurement: The Contested Count of the Disappeared in Argentina'. *Human Rights Quarterly* 16(4):676–92.

Cingranelli, David L., and David L. Richards. 1999. 'Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights'. *International Studies Quarterly* 43(2):407–17.

——. 2010. 'The Cingranelli and Richards (CIRI) Human Rights Data Project'. *Human Rights Quarterly* 32(2):401–24.

——. 2012a. 'The Cingranelli-Richards (CIRI) Human Rights Data Project Coding Manual Version 2008.3.13', available at http://ciri.binghamton.edu/documentation/ciri_coding_guide.pdf.

[22] Many scholars have argued that event-based data (rather than the standards-based PTS) are a more appropriate operationalization of human rights respect (e.g., Davenport 1995; Lopez and Stohl 1992). See Poe (2004) for a short review of the literature that critiques the cross-sectional comparison of event-based data.

Cingranelli, David L., and David L Richards. 2012b. 'The Cingranelli-Richards Human Rights Dataset Version 2008.03.12', available at http://www.humanrightsdata.org.

Clark, Ann Marie, and Kathryn Sikkink. Forthcoming. 'Information Effects and Human Rights Data: Is the Good News about Increased Human Rights Information Bad News for Human Rights Measures?' *Human Rights Quarterly*.

Claude, Richard P., and Thomas B. Jabine. 1986. 'Symposium: Statistical Issues in the Field of Human Rights'. *Human Rights Quarterly* 8(4):551–66.

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

Conrad, Courtenay R., and Jacqueline H.R. Demeritt. 2011. 'Options in the Arsenal: Are Repressive Tactics Substitutes or Complements?' Working Paper.

Davenport, Christian. 1995. 'Multi-Dimensional Threat Perception and State Repression: An Inquiry into Why States Apply Negative Sanctions'. *American Journal of Political Science* 39(3):683–713.

Desposato, Scott W. 2006. 'The Impact of Electoral Rules on Legislative Parties: Lessons from the Brazilian Senate and Chamber of Deputies'. *Journal of Politics* 68(4):1018–30.

Donnelly, Jack, and Rhoda E. Howard. 1988. 'Assessing National Human Rights Performance: A Theoretical Framework'. *Human Rights Quarterly* 10(2):214–48.

Duvall, Raymond D., and Michael Stohl. 1983. 'Governance by Terror'. In *The Politics of Terrorism*, edited by Michael Stohl, 179–219. New York: Marcel Dekker.

Fariss, Christopher J., and Keith Schnakenberg. 2013. 'Measuring Mutual Dependence Between State Repressive Actions'. *Journal of Conflict Resolution* forthcoming.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 2003. *Bayesian Data Analysis, Second Edition*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, Andrew, and Donald B. Rubin. 1992. 'Inference from Iterative Simulation using Multiple Sequences'. *Statistical Science* 7:457–511.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press.

Geweke, John. 1992. Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments. In *Bayesian Statistics 4*, edited by J. M. Bernardo, J. Berger, A. P. Dawid and J. F. M. Smith, 169–93. Oxford: Oxford University Press.

Gill, Jeff. 2007. *Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences), Second Edition*. Boca Raton, FL: Chapman and Hall/CRC.

Goodman, Ryan, and Derek Jinks. 2003. 'Measuring the Effects of Human Rights Treaties'. *European Journal of International Law* 14(1):171–83.

Heidelberger, Philip, and Peter D. Welch. 1981. 'A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations'. *Communications of the ACM* 24(4):233–45.

——. 1983. 'Simulation Run Length Control in the Presence of an Initial Transient'. *Operations Research* 31(6):1109–44.

Jabine, Thomas B., and Richard P. Claude, eds. 1992. *Human Rights and Statistics: Getting the Record Straight*. Philadelphia: University of Pennsylvania Press.

Jackman, Simon. 2008. 'Measurement'. In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press.

——. 2009. *Bayesian Analysis for the Social Sciences*. New York: John Wiley and Sons.

Kass, Robert E., and Adrian E. Raftery. 1995. 'Bayes Factors'. *Journal of the American Statistical Association* 90(430):773–95.

Keck, Margaret, and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

Landman, Todd. 2005. 'The Political Science of Human Rights'. *British Journal of Political Science* 35(3):549–72.

Landman, Todd, and Edzia Carvalho. 2009. *Measuring Human Rights*. London: Routledge.
Lauderdale, Benjamin E. 2010. 'Unpredictable Voters in Ideal Point Estimation'. *Political Analysis* 18(2):151–71.
Lopez, George A., and Michael Stohl. 1992. 'Problems of Concept and Measurement in the Study of Human Rights'. In *Human Rights and Statistics: Getting the Record Straight*, edited by Thomas B. Jabine and Richard P. Claude. Philadelphia: University of Pennsylvania Press.
Martin, Andrew D., and Keven M. Quinn. 2002. 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999'. *Political Analysis* 10(2): 134–53.
McCormick, James M., and Neil J. Mitchell. 1997. 'Human Rights Violations, Umbrella Concepts, and Empirical Analysis'. *World Politics* 49(4):510–25.
Meernik, James D., Rosa Aloisi, Marsha Sowell, and Angela Nichols. 2012. 'The Impact of Human Rights Organizations on Naming and Shaming Campaigns'. *Journal of Conflict Resolution* 56(2):233–56.
Mislevy, Robert. 1991. 'Randomization-based Inference about Latent Variables from Complex Samples'. *Psychometrika* 56(2):177–96.
Murdie, Amanda, and David R. Davis. 2012. 'Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs'. *International Studies Quarterly* 56(1):1–16.
Pemstein, Daniel, Stephen A. Meserve, and James Melton. 2010. 'Democratic Compromise: A Latent Variable Analysis to Ten Measure of Regime Type'. *Political Analysis* 18(4):426–49.
Piazza, James A., and James Igoe Walsh. 2009. 'Transnational Terror and Human Rights'. *International Studies Quarterly* 53(1):125–48.
Plummer, Martyn. 2010. 'JAGS (Just Another Gibbs Sampler) 1.0.3 Universal', available from http://www-fis.iarc.fr/ martyn/software/jags/.
Poe, Steven C. 1990. 'Human Rights and US Foreign Aid: A Review of Quantitative Studies and Suggestions for Future Research'. *Human Rights Quarterly* 12(4):499–512.
——. 1991. 'U.S. Economic Aid Allocation: The Quest for Cumulation'. *International Interactions* 16(4):295–316.
——. 2004. 'The Decision to Repress: An Integrative Theoretical Approach to the Research on Human Rights and Repression'. In *Understanding Human Rights Violations: New Systematic Studies*, edited by S. Carey and S. Poe, 16–42. Aldershot: Ashgate.
Poe, Steven C., Sabine C. Carey, and Tanya C. Vazquez. 2001. 'How are These Pictures Different? A Quantitative Comparison of the US State Department and Amnesty International Human Rights Reports, 1976–1995'. *Human Rights Quarterly* 23(3):650–77.
Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge, UK: Cambridge University Press.
Poole, Keith T., and Howard. Rosenthal 1991. 'Patterns of Congressional Voting'. *American Journal of Political Science* 35(1):228–78.
——. 1997. *A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.
Quinn, Keven M. 2004. 'Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses'. *Political Analysis* 12(4):338–53.
R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, available from http://www.R-project.org.
Richards, David, Ronald Gelleny, and David Sacko. 2001. 'Money With A Mean Streak? Foreign Economic Penetration and Government Respect for Human Rights in Developing Countries'. *International Studies Quarterly* 45(2):219–39.
Rosas, Guillermo. 2009. 'Dynamic Latent Trait Models: An Application to Latin American Banking Crises'. *Electoral Studies* 28:375–87.
Rosas, Guillermo, and Yael Shomer. 2008. 'Models of Nonresponse in Legislative Politics'. *Legislative Studies Quarterly* 33(4):573–601.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Schenker, Nathaniel, and Jane F. Gentleman. 2001. 'On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals'. *The American Statistician* 55(3):182–6.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. 2002. 'Bayesian Measures of Model Complexity and Fit'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4):583–639.

Stohl, Michael, David Carleton, George Lopez, and Stephen Samuels. 1986. 'State Violation of Human Rights: Issues and Problems of Measurement'. *Human Rights Quarterly* 8(4):592–606.

Treier, Shawn, and Simon Jackman. 2008. 'Democracy as a Latent Variable'. *American Journal of Political Science* 52(1):201–17.

Voeten, Erik. 2000. 'Clashes in the Assembly'. *International Organization* 54(2):185–215.

Wood, Reed M., and Mark Gibney. 2010. 'The Political Terror Scale (PTS): A Re-introduction and Comparison'. *Human Rights Quarterly* 32(2):367–400.

APPENDIX

The JAGS code that estimates the latent variables, displayed below, was run using Martyn Plummer's JAGS software (Plummer 2010). All other estimations were run in R (R Development Core Team 2011) using the statistical packages coda, Rjags and R2jags. All data and code used in this analysis are publicly available at a Dataverse repository maintained by the authors.

DO-IRT JAGS Code for Physical Integrity

```
model {
  for(i in 1:n){# n is the number of obs
    for(item in 1:4){
      logit(Z[i, item, 1]) <- alpha[item, 1] - beta[item]*x[i]
      logit(Z[i, item, 2]) <- alpha[item, 2] - beta[item]*x[i]
      Pi[i, item, 1] <- Z[i, item, 1]
      Pi[i, item, 2] <- Z[i, item, 2] - Z[i, item, 1]
      Pi[i, item, 3] <- 1 - Z[i, item, 2]
      y[i, item] ~ dcat(Pi[i, item, 1:3])
    }
    x[i] <- mu[country[i], year[i]]
  }
    sigma ~ dunif(0,1)
    kappa <- pow(sigma, -1)
  for(c in 1:n.country){
    mu[c, 1] ~ dnorm(0, 1)
    for(t in 2:n.year) {#n.year is number of years
      mu[c, t] ~ dnorm(mu[c, t-1], kappa)
    }
  }
  for(j in 1:4){
    beta[j] ~ dgamma(4, 3)
    alpha0[j, 1] ~ dnorm(0, .25)
    alpha0[j, 2] ~ dnorm(0, .25)
    alpha[j, 1:2] <- sort(alpha0[j, 1:2])
  }
}
```

DO-IRT JAGS Code for Empowerment

```
model{
  for(i in 1:n){# n is the number of obs
    for(item in 1:7){
      logit(Z[i, item, 1]) <- alpha[item, 1] - beta[item]*x[i]
      logit(Z[i, item, 2]) <- alpha[item, 2] - beta[item]*x[i]
      Pi[i, item, 1] <- Z[i, item, 1]
      Pi[i, item, 2] <- Z[i, item, 2] - Z[i, item, 1]
      Pi[i, item, 3] <- 1 - Z[i, item, 2]
      y[i, item] ~ dcat(Pi[i, item, 1:3])
    }
    x[i] <- mu[country[i], year[i]]
  }
      sigma ~ dunif(0,1)
      kappa <- pow(sigma, −1)
    for(c in 1:n.country){
      mu[c, 1] ~ dnorm(0, 1)
      for(t in 2:n.year) #n.year is number of years
      mu[c, t] ~ dnorm(mu[c, t-1], kappa)
    }
  }
  for(j in 1:7){
    beta[j] ~ dgamma(4, 3)
    alpha0[j, 1] ~ dnorm(0, .25)
    alpha0[j, 2] ~ dnorm(0, .25)
    alpha[j, 1:2] <- sort(alpha0[j, 1:2])
  }
}
```

O-IRT JAGS Code for Physical Integrity

```
model{
  for(i in 1:n){# n is the number of obs
    for(item in 1:4){
      logit(Z[i, item, 1]) <- alpha[item, 1] - beta[item]*x[i]
      logit(Z[i, item, 2]) <- alpha[item, 2] - beta[item]*x[i]
      Pi[i, item, 1] <- Z[i, item, 1]
      Pi[i, item, 2] <- Z[i, item, 2] - Z[i, item, 1]
      Pi[i, item, 3] <- 1 - Z[i, item, 2]
      y[i, item] ~ dcat(Pi[i, item, 1:3])
    }
    x[i] <- mu[country[i], year[i]]
  }
  for(c in 1:n.country){
    mu[c, 1] ~ dnorm(0, 1)
    for(t in 2:n.year) {#n.year is number of years
      mu[c, t] ~ dnorm(0, 1)
    }
  }
  For(j in 1:4)
    beta[j] ~ dgamma(4, 3)
    alpha0[j, 1] ~ dnorm(0, .25)
    alpha0[j, 2] ~ dnorm(0, .25)
    alpha[j, 1:2] <- sort(alpha0[j, 1:2])
  }
}
```

O-IRT JAGS Code for Empowerment

```
model{
  for(i in 1:n){# n is the number of obs
    for(item in 1:7){
      logit(Z[i, item, 1]) <- alpha[item, 1] - beta[item]*x[i]
      logit(Z[i, item, 2]) <- alpha[item, 2] - beta[item]*x[i]
      Pi[i, item, 1] <- Z[i, item, 1]
      Pi[i, item, 2] <- Z[i, item, 2] - Z[i, item, 1]
      Pi[i, item, 3] <- 1 - Z[i, item, 2]
      y[i, item] ~ dcat(Pi[i, item, 1:3])
    }
    x[i] <- mu[country[i], year[i]]
  }
  for(c in 1:n.country){
    mu[c, 1] ~ dnorm(0, 1)
    for(t in 2:n.year) #n.year is number of years
      mu[c, t] ~ dnorm(0, 1)
    }
  }
  for(j in 1:7){
    beta[j] ~ dgamma(4, 3)
    alpha0[j, 1] ~ dnorm(0, .25)
    alpha0[j, 2] ~ dnorm(0, .25)
    Alpha[j, 1:2] <- sort(alpha0[j, 1:2])
  }
}
```

ADDITIONAL FIGURES

Here we present cross-sectional comparisons of estimates from the DO-IRT physical integrity model and the DO-IRT empowerment model.

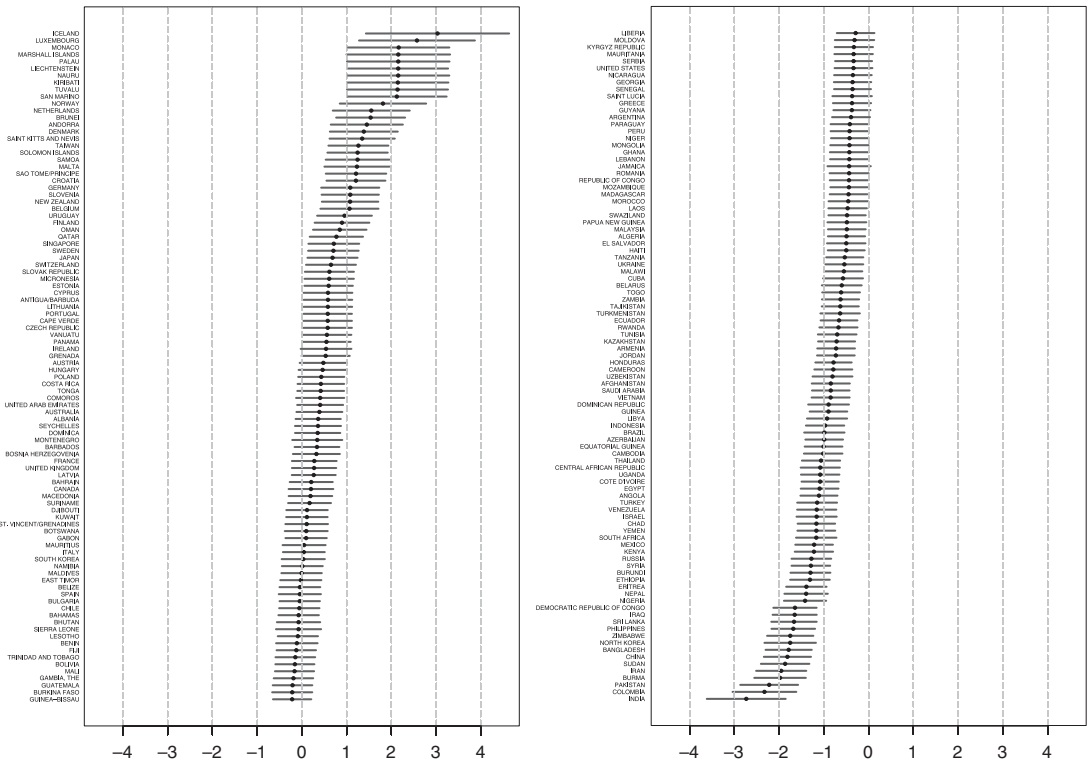

Fig. A1. DO-IRT physical integrity latent variable estimates for 192 states in 2008
Note: Dots are point estimates (posterior means) and lines are 95 percent credible intervals.

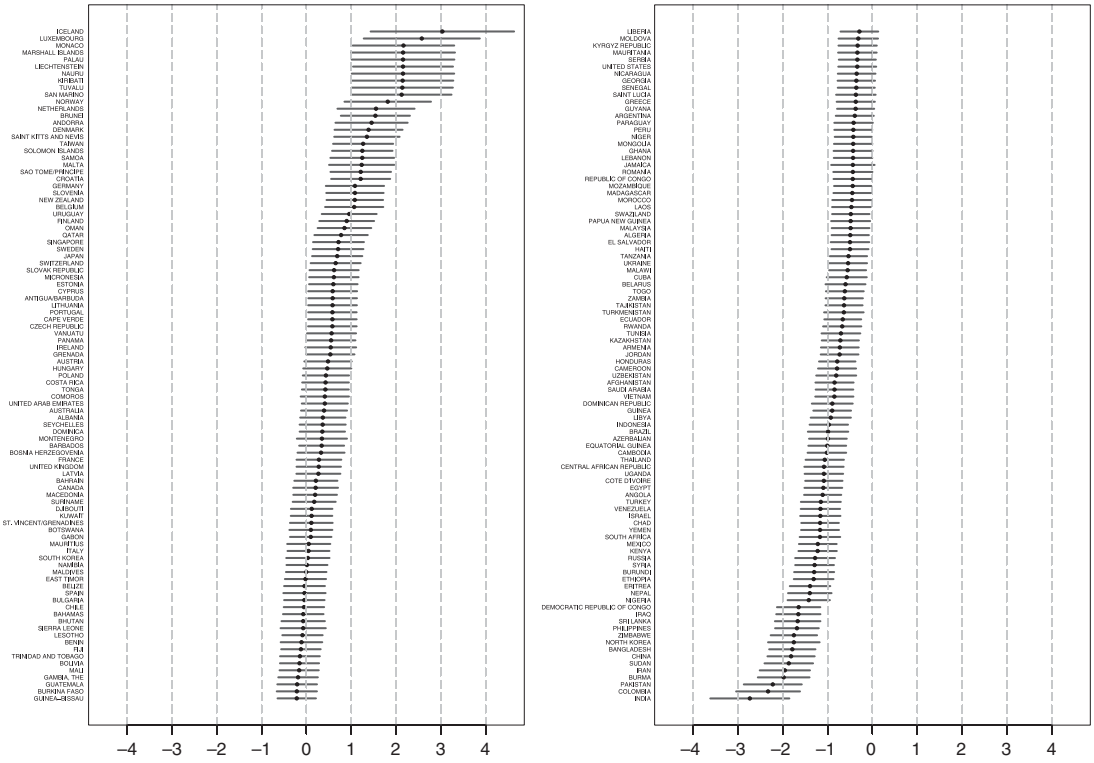

Fig. A2. DO-IRT empowerment latent variables estimates for 192 states in 2008
Note: Dots are point estimates (posterior means) and lines are 95 percent credible intervals.