doi:10.1017/S0003055414000070

Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability

CHRISTOPHER J. FARISS *Pennsylvania State University*

According to indicators of political repression currently used by scholars, human rights practices have not improved over the past 35 years, despite the spread of human rights norms, better monitoring, and the increasing prevalence of electoral democracy. I argue that this empirical pattern is not an indication of stagnating human rights practices. Instead, it reflects a systematic change in the way monitors, like Amnesty International and the U.S. State Department, encounter and interpret information about abuses. The standard of accountability used to assess state behaviors becomes more stringent as monitors look harder for abuse, look in more places for abuse, and classify more acts as abuse. In this article, I present a new, theoretically informed measurement model, which generates unbiased estimates of repression using existing data. I then show that respect for human rights has improved over time and that the relationship between human rights respect and ratification of the UN Convention Against Torture is positive, which contradicts findings from existing research.

INTRODUCTION

Have levels of political repression changed? "Repression" or violations of "physical integrity rights" include arrests and political imprisonment, beatings and torture, extrajudicial executions, mass killings, and disappearances, all of which are practices used by political authorities against those under their jurisdiction.[1] This question is important because current indicators of political repression imply that human rights practices have been essentially constant over the last 35 years (see Figure 1), despite the spread of human rights norms, better monitoring by private and public agencies, and the increasing prevalence of electoral democracy. While some theorists take issue with this empirical pattern and the data used to support it,[2] hundreds of studies rely on these indicators to analyze the determinants of repression[3] and the effects of international institutions on human rights treaty compliance.[4]

The data generated by several coding projects have come to dominate the quantitative study of human rights, yet contested empirical patterns have emerged regarding the overall trend in human rights respect and the relationship between the state behaviors and domestic and international institutions, such as the UN Convention Against Torture. I argue that the pattern of constant abuse found in data derived from human rights reports is not an indication of stagnating human rights practices. Instead, it reflects a systematic change in the way monitoring agencies, like Amnesty International and the U.S. State Department, encounter and interpret information about human rights abuses. Over time, this process has led to what I call a *changing standard of accountability*. As a consequence of this change, human rights reports have become increasingly stringent assessments of state behaviors. This change occurs because (1) government authorities have an incentive to hide the use of these policy tools and (2) observers and activists use countervailing strategies in order to reveal, understand, and ultimately change repressive practices for the better. This interaction between state actors and observers, both academic and activist, affects the production of information used by researchers to quantify repressive behaviors.[5]

Christopher J. Fariss is Assistant Professor, Department of Political Science, Pennsylvania State University (cjf20@psu.edu; cjf0006@gmail.com).

This paper was presented at the 2013 meeting of the Midwest Political Science Association and the 2013 meeting of the Western Political Science Association, the Department of Government, College of William and Mary, October, 24, 2012, the 2012 Human Nature Group Retreat, and two graduate human rights seminars led by Christian Davenport and Will Moore, respectively. I would like to thank the participants at these talks and also Megan Becker, Sam Bell, Rob Bond, Chad Clay, Geoff Dancy, Jesse Driscoll, James Fowler, Erik Gartzke, Micah Gell-Redman, Benjamin Graham, Dimitar Gueorguiev, Alex Hughes, Jason Jones, Miles Kahler, David Lake, Brad LeVeck, Yon Lupu, Kelly Matush, Blake McMahon, Jonathan Markowitz, Jamie Mayerfeld, Amanda Murdie, Maya Oren, Kai Ostwald, Keith Schnakenberg, Paul Schuler, Brice Semmens, Jaime Settle, Neil Visalvanich, Reed Wood, Thorin Wright, four anonymous reviewers, and the editors at the *American Political Science Review* for many helpful comments and suggestions. The estimates from this paper along with the code necessary to implement the models in JAGS and R are publicly available at: http://dvn.iq.harvard.edu/dvn/dv/HumanRightsScores.

[1] This definition is a modified version of one from Goldstein (1978). His definition includes censorship, which I exclude in order to focus exclusively on physical integrity violations, which are the most commonly analyzed rights.

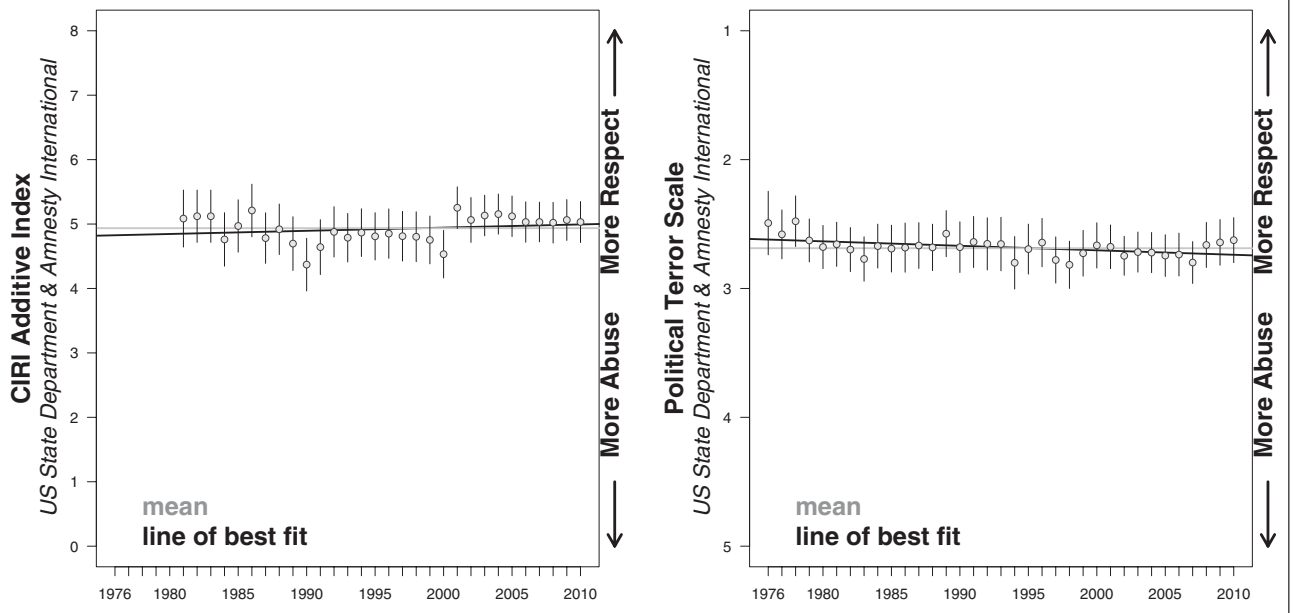
[2] See the discussion in Clark and Sikkink (2013) and Goodman and Jinks (2003).

[3] See, for example, the research by Apodaca (2001), Bell, Clay and Murdie (2012), Bueno De Mesquita et al. (2005), Cingranelli and Filippov (2010), Conrad and Moore (2010), Davenport (1995), Davenport (2010), Davenport and Armstrong (2004), Demeritt and Young (2013), Fariss and Schnakenberg (2013), Nordås and Davenport (2013), Poe and Tate (1994), Poe, Tate and Keith (1999), Wood (2008), and Zanger (2000). Also see the reviews by Davenport (2007*a*) and Poe (2004).

[4] See, for example, the research by Conrad and Ritter (2013), Dancy and Sikkink (2012), Hathaway (2002), Hafner-Burton and Tsutsui (2005), Hafner-Burton and Tsutsui (2007), Hill (2010), Keith (1999), Keith, Tate and Poe (2009), Lupu (2013*a*), Lupu (2013*b*), Neumayer (2005), and Simmons (2009).

[5] Though human rights theorists are aware of this issue (Brysk 1994; Clark 2001; Goodman and Jinks 2003; Hill, Moore and Mukherjee

FIGURE 1. Yearly Mean and 95% Confidence Intervals for the Estimated Level of Repression Using the CIRI Additive Index (left), and the Political Terror Scale Index (right)



Notes: Each series is based on the human rights reports from the U.S. State Department and Amnesty International. Note that the averages for the Political Terror Scale estimates are based on two scales coded independently, one from the U.S. State Department reports and one from the Amnesty International reports. Similar figures for the individual PTS variables are displayed in Appendix D.

In this article, I present results for a new view of repression: *physical integrity practices have improved over time*. To support my claim, I compare an existing dynamic ordinal item response theory model (Schnakenberg and Fariss 2014), which I call the *constant standard model*, to a new extension of this model, which I call the *dynamic standard model*.[6] The constant standard model, like other existing models of repression (e.g., the CIRI Additive index, the Political Terror Scale index, the Hathaway torture index, and the latent variable model developed by Schnakenberg and Fariss), implicitly assumes that the standard of accountability does not change over time. The dynamic standard model relaxes this assumption. Thus, by comparing the information derived from these models, I am able to demonstrate that unobserved changes to the standard of accountability explain why average levels of repression have appeared to remain unchanged as all existing models of human rights suggest.

In sum, for the constant standard model to be more consistent with reality and for this same pattern to obtain, Amnesty International and the U.S. State Department would need to produce human rights reports consistently from year to year *and* the producers of the event-based data (introduced below) would need to use a less and less stringent definition of repression in the assessment of repressive events over time. This is unlikely because the event-based variables I introduce below are consistently updated as new information about the specific events becomes available. In addition to periodic updating, the producers of these event-based variables are focused on the extreme end of the repression spectrum (e.g., genocide, politicide, mass repression). Both of these features suggest that the event-based data are a valid representation of the historical record to date. The event-based data therefore act as a consistent baseline by which to compare the levels of the standards-based variables, which are produced in a specific historical context and never updated. I provide more detailed information about this point below.

The solution proposed here may be applied more broadly to data derived from other documents that might vary systematically. Issues of comparability are of increasing concern for social scientists—not just human rights scholars—interested in comparing data derived from human rights documents over time. This is because recent advances in computational throughput, digital storage, and automated content methods have made the analysis of large scale corpuses of primary source documents cost effective and increasingly popular. Though this article is not about automated content analysis per se, it is concerned with the comparison of coded documents over time. As the tools of automated content analysis become more popular and accepted in the mainstream of political science and academia

2013; Keck and Sikkink 1998), this is the first project that systematically incorporates it into a measurement model of repression.

[6] The dynamic standard model formalizes the relationship between the unmeasured standard of accountability and observed levels of repression measured by several existing data sources. Note that both of the models compared in this article are dynamic with respect to the estimated country-year latent variable but not the item difficulty parameters. It is this difference that allows me to estimate changes to the standard of accountability over time. I describe these model features in the methods section below.

generally, it will be all the more important to determine if the primary source documents selected for analysis are comparable. That is, have the documents systematically changed, thus biasing the resulting codings? If so, what are the solutions to such issues? In this article, I offer an applied solution for this problem in the context of hand-coded human rights documents.

Overall, I make several important contributions: First, I develop a theory of the standard of accountability and a new measurement model that accounts for it. Second, though the primary motivation of this article is substantive, I make two methodological contributions, as the measurement model I develop is the first in the political science literature to estimate time-varying item-difficulty cut points for some variables (i.e., the repression outcome variables).[7] More broadly, I highlight the importance of determining if primary source political texts selected for analysis are comparable. This measurement issue exists independent of whether the documents are coded by experts or using the increasingly popular automated content analysis techniques. Third, I introduce and make publicly available new unbiased country-year estimates of repression that cover the time period beginning in 1949 and ending in 2010 ($n = 9267$). The resulting data are the most comprehensive estimates of repression that currently exist and will help academics and policy makers begin to reassess what has become common knowledge in the human rights literature.[8] Fourth, I provide empirical evidence that human rights practices have improved over time, which contradicts the patterns found in existing data. Fifth, I illustrate the substantive importance of the results to international relations theory by showing that the relationship between human rights respect and ratification of the UN Convention Against Torture is positive, which contradicts controversial findings from existing research. Finally, I offer suggestions for how to correct for temporal bias in standard models of repression, which will allow researchers to continue to analyze the original ordered human rights variables coded directly from human rights country reports.

WHY THE STANDARD OF ACCOUNTABILITY CHANGES OVER TIME

The standard of accountability, which I define as the set of expectations that monitoring agencies use to hold states responsible for repressive actions, has not been systematically addressed because much of the data measuring repression are derived from the same primary sources (Cingranelli and Richards 1999, 2012*a*, *b*; Gibney, Cornett and Wood 2012; Hathaway 2002). The documents used to measure repression are *The Country Reports on Human Rights Practices* published annually by the U.S. State Department and *The State of the World's Human Rights* report published annually by Amnesty International. The information captured in these documents will bias assessments of repression over time if changes to these documents are not also taken into account. I argue that the standard of accountability changes over time because of the tactics used by reporting agencies to (1) gather accurate information about credible allegations of repression, (2) broaden the coverage of information gathering campaigns with the help of other NGOs, and (3) continually press governments to reform through naming and shaming campaigns, even after real reforms are implemented to reduce more egregious rights violations by those governments. In the language of research design, *instrumentation bias* occurs if the measurement tool used to assess a behavior changes over time (Trochim and Donnelly 2008).

The standard of accountability changes due to a combination the three tactics, described above, which I will refer to as (1) information, (2) access, and (3) classification. These tactics make up the strategies used by observers and activists interested in revealing, understanding, and ultimately changing repressive practices for the better. First, improvements in the quality and increases in the quantity of information have led to more accurate assessments of the conditions in each country over time. Second, access to countries by NGOs, like Amnesty International and Human Rights First (formerly the Lawyer's Committee for Human Rights), which seek to collect and disseminate accurate information about repression allegations and practices, has increased as these organizations grow and cooperate with one another. Third, changes in the subjective views of what constitutes a "good" human rights record held by analysts at the monitoring agencies are anchored by the status quo, which improves as the global average of rights respect improves. That is, monitoring agencies continually press governments to institute new reforms, even after real reforms are implemented to reduce more egregious rights violations like extrajudicial killings and disappearances. Thus, the set of expectations that monitoring agencies use to assess and document state behaviors changes over time as these monitors look harder for abuse, look in more places for abuse, and classify more acts as abuse. In the remainder of this section, I discuss each of these tactics in more detail and provide examples illustrating how monitoring agencies have used these tactics to respond to changes in state behaviors over time.

Human rights theorists recognize that the information used to assess government behaviors may change over time and that this could mask underlying improvements in human rights practices.[9] Keck and Sikkink (1998) attribute this change to an "information paradox." The paradox occurs when an increase in information leads to difficulties in assessing the efficacy of advocacy campaigns over time because of the very success in collecting and aggregating accounts of repressive actions in the first place. Clark and Sikkink

[7] These model parameters measure changes to the standard of accountability over time.

[8] Such empirical relationships include, but are not limited to, the effect of sanctions, naming and shaming campaigns, foreign aid, and domestic and international legal institutions on the level of respect for human rights.

[9] See, for example, the research by Bollen (1986), Brysk (1994), Clark (2001), Goodman and Jinks (2003), and Keck and Sikkink (1998).

(2013) coin a similar term—"human rights information paradox"—to describe this issue as it relates to human rights abuses specifically. As a result of this paradox, the global human rights situation may appear to have worsened over time because there is simply an increasing amount of information with which to assess human rights practices. Moreover, Innes de Neufville (1986) argues that the quality of the human rights reports produced by the U.S. State Department increased because of changes to the reporting requirements, which "altered practices and norms within the Department of State and created an arena for public evaluation of the information" (682). Improvements in the reports are corroborated by yearly critiques published by the Lawyers Committee for Human Rights and a quantitative analysis by Poe, Carey and Vazquez (2001). Thus, monitoring agencies look harder for information about human rights abuse over time.

In addition to the quality and quantity of information, access to government documentation, witnesses, victims, prison sites, and other areas are important for assessing state behaviors. Both Amnesty International and the U.S. State Department rely on reports from other NGOs that collect and disseminate information about human rights abuses within states. The number and effectiveness of these actors have increased over time, especially since the end of the Cold War.[10] Moreover, as Hill, Moore and Mukherjee (2013) argue, increasing numbers of domestic NGOs generate more credible signals about government abuse, which are then used by Amnesty International and, by extension, the U.S. State Department in the production of human rights reports. Monitors are therefore capable of looking in more places for abuse, as their numbers increase and their collaborative networks develop and expand over time (Keck and Sikkink 1998).

Monitoring agencies are also increasingly sensitive to the various kinds of ill-treatment that previously fell short of abuse but that still constitute violations of human rights. Therefore, monitors such as Amnesty International are continually pressing for additional reforms through naming and shaming campaigns, even as more egregious violations by state actors cease to occur. As Sikkink notes, these monitoring agencies and others "have expanded their focus over time from a narrow concentration on direct government responsibility for the death, disappearance, and imprisonment of political opponents to a wider range of rights, including the right of people to be free from police brutality and the excessive use of lethal force" (2011, 159). Moreover, there is specific evidence from case law of a rising standard of acceptable treatment, whereby more acts come to be classified as inhuman treatment or torture over time. For example, the European Court of Human Rights, in Selmouni v. France (1999), "consider certain acts which were classified in the past as 'inhuman and degrading treatment' as opposed to 'torture' could be classified differently in future." That is, acts by state agents that might have previously been classified within the less severe category of ill-treatment and degrading punishment might now be classified as torture. The court states further "that the increasingly high standard being required in the area of the protection of human rights and fundamental liberties correspondingly and inevitably requires greater firmness in assessing breaches of the fundamental values of democratic societies."[11] As this example illustrates, human rights monitors are increasingly stringent in their assessment of state behaviors precisely because they are classifying more acts as abuse over time and because they are continually pressing for additional reforms from states even as those very states limit and eventually cease the use of more egregious tactics.

Overall, the standard of accountability becomes more stringent as the U.S. State Department and Amnesty International look harder for abuse, look in more places for abuse, and classify more acts as abuse. For example, Amnesty International expanded its strategy over time as it responded to developments in the repressive behaviors used by states.[12] The initial focus of Amnesty International on political prisoners during the 1960s and 1970s precluded the reporting of extrajudicial killings that took place outside of prisons (Clark, 2001, chap. 5). Also during the 1960s and 1970s, state agents in Guatemala frequently disappeared opposition members, yet Amnesty International did not document these policies until 1976, because these actions were not initially a policy tool of concern (Clark, 2001, chap. 4).

For another illustration, consider the text from the torture section contained in the State Department human rights reports on Guatemala for the years 1981, 1991, and 2001. The document as a whole and the torture section in particular both provide more detailed information in later years about the government agencies committing human rights violations and the groups experiencing those violations when compared to earlier years. The word count of the document and torture section both dramatically increase in length as displayed in Table 1 as well. Consider just one of the human rights variables discussed in more detail below: the differences between the coding of "frequent torture" on the CIRI Torture scale in 2001 relative to the less severe coding in 1991 could be a function of the amount of information and the specificity of the information included in the reports in the different years. It might also be due to increasing coverage of NGOs working in this country. It could also be because of the

[10] See discussions in Hopgood (2006), Hill, Moore and Mukherjee (2013), Korey (2001), Keck and Sikkink (1998), Lake and Wong (2009), Murdie and Bhasin (2011), Murdie and Davis (2012), and Wong (2012).

[11] Selmouni v. France, 25803/94, Council of Europe: European Court of Human Rights, 28 July 1999, available at: http://www.unhcr.org/refworld/docid/3ae6b70210.html. I would like to thank Jamie Mayerfeld for pointing me towards this example and for helping me clarify this point.

[12] See Clark (2001) for a discussion of the developments in the strategy used by Amnesty International in response to changes in repressive behaviors. Berman and Clark (1982) provides an example of how political authorities in the Philippines began to disappear political opponents to avoid public scrutiny of other human rights violations.

TABLE 1. Changing Information Content in Three Human Rights Reports

Year	Torture Section Word Count	Full Document Word Count	CIRI Torture Coding
1981	329	3,930	0 (frequent)
1991	562	5,768	1 (some)
2001	3,669	32,064	0 (frequent)

increasingly stringent standard (the rising standard of acceptable treatment) being required in the area of the protection of human rights and fundamental liberties or the continued pressure for reform over time, which are captured in the text of these documents. As these cases suggest, the reports published today represent a broader and more detailed view of the human rights practices than reports published in previous years (See Appendix C for examples of the text from the torture section for these country-years).[13]

The standard of accountability has increased due to at least one if not all of the mechanisms outlined above. Thus, as the theory suggests, the set of expectations that monitoring agencies use to hold states responsible for repressive actions changes over time. Unfortunately for scholars interested in these changes, the standard of accountability is not observable in human rights reports and is therefore impossible to directly measure. To make matters more complicated, alternative sources of information that were once highly cited are now largely forgotten and out of date (Harff and Gurr 1988; Rummel 1994*a*, *b*, 1995; Taylor and Jodice 1983). Contemporary alternatives often cover shorter periods of time (Conrad, Haglund and Moore 2013; Eck and Hultman 2007), are not up to date (Conrad, Haglund and Moore 2013; Hathaway 2002), or still rely on the same standards-based human rights reports (Hathaway 2002). All of these issues make the systematic comparison of results from different data sources difficult, which leaves the problem of instrumentation bias acknowledged but unaddressed in the literature. Fortunately, these issues can now be overcome thanks to the wide availability of computational tools capable of linking diverse sources of data in theoretically meaningful ways. The latent variable models I describe below are capable of (1) bringing together diverse sources of information, (2) assessing the relative quality of the information included, and (3) quantifying the certainty of the estimates of repression that are generated from the models. These models allow me to test for the influence of the standard of accountability by comparing the new model in which the probability of documenting a repressive action changes over time (dynamic standard model) to the existing model in which this probability does not change (constant standard model). In the next section, I introduce and make a theoretical distinction between standards-based data and event-based data. I then introduce the latent variable models.

[13] I am systematically exploring the differences in the quality and quantity of information in the text of human rights documents in a book-length project that builds off the insights from this article.

STANDARDS-BASED DATA AND EVENT-BASED DATA

The event-based variables (*sources*: multiple, see Table 3) introduced in this section act as a consistent baseline by which to compare the levels of the standards-based variables (*sources*: Amnesty International and the U.S. State Department; see Table 2), which are produced in a specific historical context and never updated. This is because the event-based variables included in this article are consistently updated as new information about the specific events becomes available. In addition to periodic updating, the producers of these event-based variables are focused on the extreme end of the repression spectrum (e.g., genocide, politicide, mass repression). This focus makes identifying these events much easier than more difficult to observe behaviors, such as torture. Both of these features suggest that the event-based data included in this article are a valid representation of the historical record to date, which acts as a baseline for the latent variable model described in detail in the next section.

Before discussing the theoretical differences between standards-based data and event-based data in more detail, the reader should keep in mind that all of the variables included in the two competing latent variable models are the same and that they are each operationalized to capture one or more of the repressive behaviors identified in the definition of repression used throughout this article. Recall that the definition of "repression" or violations of "physical integrity rights" and sometimes called "state sanctioned terror" includes arrests and political imprisonment, beatings and torture, extrajudicial executions, mass killings, and disappearances, all of which are practices used by political authorities against those under their jurisdiction. The standards-based and event-based variable names, operationalizations, citations, and data sources are displayed in Tables 2 and 3, respectively. The temporal coverage and data type of each variable are displayed in Figure 2.

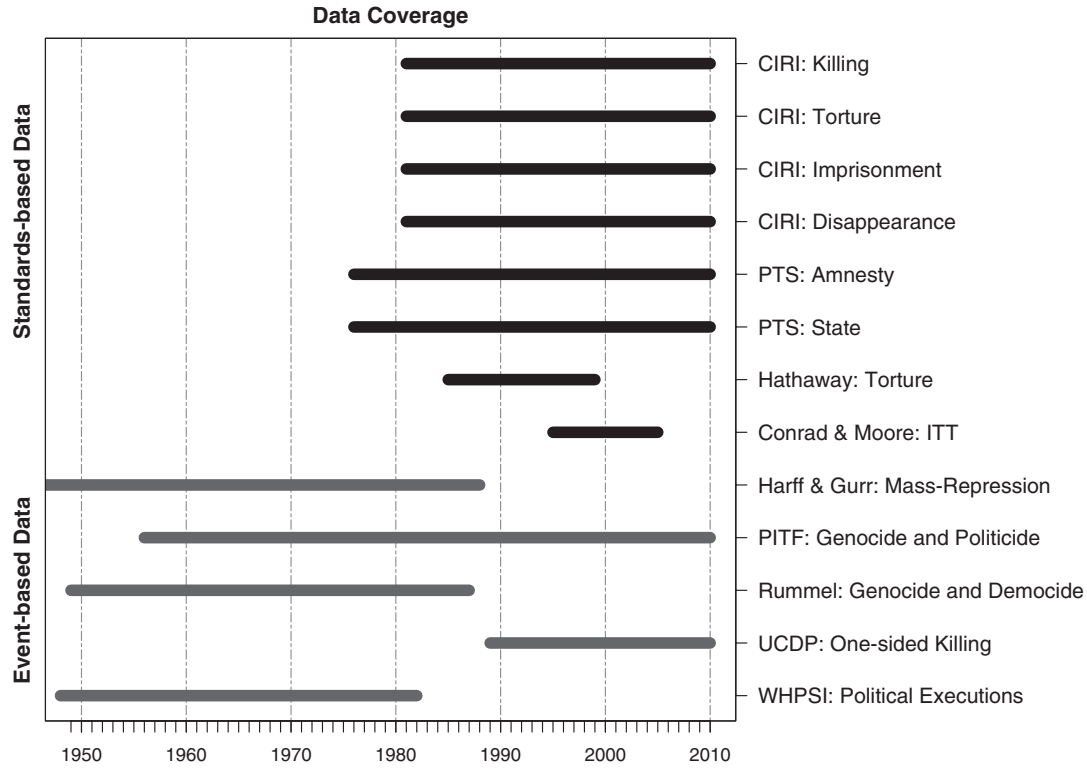
In some cases, leaders or regimes choose to attempt the complete elimination of a political group (politicide) or other group designation (genocide) (Harff 2003; Rummel 1994*b*, 1995). Massive repressive events are related to the intent of genocide in the use of mass slaughters or pogroms to eliminate substantial portions of a predetermined group but are a broader category that includes a greater number of events than those captured by the genocide or politicide definition (Harff and Gurr 1988). Though the definitions of genocide, politicide, and massive repression are complex (see Appendix B), each variable is focused on capturing instances of large-scale aggregated mass killings of individuals which occur for a variety reasons. Thus, these variables are focused on the extreme end of the repression spectrum. Relatedly, the measurement of one-sided government killing captures instances in which

TABLE 2. Standards-based Repression Data Sources

Dataset Name and Variable Description	**Dataset Citation** and Primary Source Information
CIRI Physical Integrity Data, 1981–2010 political imprisonment (ordered scale, 0–2) torture (ordered scale, 0–2) extrajudicial killing (ordered scale, 0–2) disappearance (ordered scale, 0–2)	Cingranelli and Richards (1999, 2012*a*, *b*) Amnesty International Reports[1] and State Department Reports[2] *Information in Amnesty reports takes precedence over information in State Department reports*
Hathaway Torture Data, 1985–1999 torture (ordered scale, 1–5)	Hathaway (2002) State Department Reports[1]
Ill-Treatment and Torture (ITT), 1995–2005 torture (ordered scale, 0–5)	Conrad and Moore (2011), Conrad, Haglund and Moore (2013), Amnesty International (2006) Annual Reports,[1] press releases,[1] and Urgent Action Alerts[1]
PTS Political Terror Scale, 1976–2010 Ammesty International scale (ordered scale, 1–5) State Department scale (ordered scale, 1–5)	Gibney, Cornett and Wood (2012), Gibney and Dalton (1996) Amnesty International Reports[1] State Department Reports[1]

Notes: [1]Primary source. [2]Secondary source.

FIGURE 2. Temporal Coverage and Data Type of Repression Data Sources (see Tables 2 and 3 for more information)



Note: Grey lines are event-based data; black lines are standards-based measures.

TABLE 3. Event-based Repression Data Sources

Dataset Name and Variable Description	**Dataset Citation** and Primary Source Information
Harff and Gurr Dataset, 1946–1988 massive repressive events (1 if country-year experienced event, 0 otherwise)	Harff and Gurr (1988) historical sources (see article references)[1]
Political Instability Task Force (PITF), 1956–2010 genocide and politicide (1 if country-year experienced event, 0 otherwise)	Harff (2003), Marshall, Gurr and Harff (2009) historical sources (see article references)[1] State Department Reports[2] Amnesty International Reports[2]
Rummel Dataset, 1949–1987 genocide and democide (1 if country-year experienced event, 0 otherwise)	Rummel (1994*b*, 1995), Wayman and Tago (2010) New York Times,[1] New International Yearbook,[2] Facts on File,[2] Britannica Book of the Year,[2] Deadline Data on World Affairs,[2] Keesing's Contemporary Archives[2]
UCDP One-sided Violence Dataset, 1989–2010 government killing (event count estimate) (1 if country-year experienced event, 0 otherwise)	Eck and Hultman (2007), Sundberg (2009) Reuters News,[1] BBC World Monitoring[1] Agence France Presse,[1] Xinhua News Agency,[1] Dow Jones International News,[1] UN Reports,[2] Amnesty International Reports,[2] Human Rights Watch Reports,[2] local level NGO reports (not listed)[2]
World Handbook of Political and Social Indicators WHPSI, 1948–1982 political executions (event count estimate) (1 if country-year experienced event, 0 otherwise)	Taylor and Jodice (1983) New York Times,[1] Middle East Journal,[2] Asian Recorder,[2] Archiv der Genenwart[2] African Diary,[2] Current Digest of Soviet Press[2]

Notes: [1]Primary source. [2]Secondary source.

more than 25 individuals (noncombatants) are killed, though this variable excludes extrajudical killings that occur inside a prison and combatant deaths that occur during civil conflicts (Eck and Hultman 2007). Extrajudicial killing more generally is captured by both the political execution data (Taylor and Jodice 1983) and several of the variables derived from the human rights reports (Cingranelli and Richards 2012*a*, *b*; Gibney, Cornett and Wood 2012). In addition to the events-based variables, the standards-based variables included in this article capture, individually or as part of an index, extrajudicial killing, in addition to torture, political imprisonment, and disappearances (Cingranelli and Richards 2012*a*, *b*; Conrad and Moore 2011; Conrad, Haglund and Moore 2013; Gibney, Cornett and Wood 2012; Hathaway 2002).[14] In sum, all of these variables capture information about the use of "repression" or violations of "physical integrity rights."

If the standard of accountability has increased over time, then comparisons of data derived from standards-based documents will be biased over time because of unaccounted for changes in the instrument (human rights documents) used to measure behavioral change (levels of repression). Temporal comparisons of categorical measures derived from the standards-based data sources are problematic because the data are based on content of reports that were prepared in a specific historical context. The reports are primary source documents that are used by analysts to derive variables on repression. The issue of temporal comparability arises because the older reports are not updated or revised even if new information about specific repressive events is obtained over time or as the goals, strategic incentives, or status quo expectations of the monitoring agencies evolve. These same issues make data derived from these reports quite useful for comparing state behaviors in the same year.

Event-based data sources contrast with the standards-based data because they are based on evidence derived from a pool of regularly augmented primary source documents. However, it is also likely that increased access to countries over time will also lead to an increase in the accuracy of the event-count data. The producers of the event-based data are aware of this process. When new information about repressive actions becomes available from NGOs, news reports, historians, or truth commissions, these scholars

[14] Conrad and Moore (2011) are quick to point out, however, that their data are designed to capture "reporting" of torture and not actual "levels" of torture. This is the only dataset that explicitly makes this theoretical distinction.

update their data. Moreover, information from multiple sources are used to help corroborate each datum.[15] These data therefore represent the best approximation of the historical pattern of repression for a given country at each update. For example, Rummel discusses the process by which he periodically updated the event-based information used in his articles and books.[16] Similar discussions can be found in the documentation of the other event-based data. Moreover, in addition to periodic updating, the producers of these event-based variables are focused on the extreme end of the repression spectrum. This focus makes identifying these events much easier than more difficult to observe behaviors such as torture. Both of these features suggest that the event-based data are a valid representation of historical record to date and therefore capable of acting as a baseline in the latent variable model described next.

Skepticism over the comparability of event data that counts the number of repressive events in country-year observations was one of the main reasons for the movement away from event data in cross-national human rights research.[17] Standards-based variables were developed in part because of the availability and comprehensiveness of the human rights reports but also in reaction to the use of event-based data.[18] I avoid this issue for now by focusing on event data that are binary, and therefore focused on country-year events that occur at the extreme end of the repression spectrum.[19] However, this raises another practical issue: binary event data only capture extreme levels of abuse. These data have been useful for comparing broad trends but the relative specificity of the standards-based reports is another reason for their preeminence over the event-based binary data. The standards-based data have provided analysts with more behavioral categories for comparison. The models I develop below can incorporate information from multiple sources, and quantify the uncertainty of each estimate, conditional on the availability of each variable included in the model. Missing data do not lead to a loss of country-year observations but only increase the uncertainty for the estimate of a given country-year. The models can also accommodate variables measured using different scales. Appendices A and B contain additional information about the development and coding rules for these variables.

TWO COMPETING LATENT VARIABLES MODELS OF REPRESSION

The latent variable models I develop in this article are item-response theory (IRT) models. The dynamic standard model is an extension of the DO-IRT (dynamic ordinal item response theory) model developed by Schnakenberg and Fariss (2014). The constant standard model is identical to the DO-IRT model. Note that both models presented in this article are dynamic with respect to the estimated latent human rights variable. The models differ with respect to the standard of accountability. In one case the standard changes (dynamic standard model) and in one case it does not change (constant standard model).

The constant standard model in addition to all of the existing models of repression—those based on information from the annual human rights reports—implicitly assumes a constant standard of accountability over time. By comparing the estimates from the constant standard model, which makes this assumption, and the dynamic standard model, which relaxes this assumption, I am able to test the hypothesis that an increase in the standard of accountability—the probability of observing and therefore coding a repressive outcome (as modeled by time-varying item cut points)—increases over time for the repression variables derived from the human rights reports.

Schnakenberg and Fariss (2014) model the latent respect for human rights for a country in a particular year as dependent on the value for the same country in the previous year. These authors demonstrate that the dynamic latent variable model fits the CIRI human rights data (Cingranelli and Richards 2012*a*, *b*) substantially better than a static latent variable model similar to those used in the democracy literature.[20] The latent variable models measuring democracy developed by Treier and Jackman (2008) and Pemstein, Meserve and Melton (2010) assume that the observed indicators used in the model are independent conditional on the value of the trait to be estimated, which is an overly strong assumption in the case of human rights as demonstrated by Schnakenberg and Fariss (2014). And, though Armstrong (2011) relaxes this assumption by using a dynamic factor analytic model to analyze the Freedom House Indicators, he models the observed indicators as interval response variables instead of ordered categories and provides no evidence that the model performs better than any alternative parameterizations.[21] As I demonstrate below, model

[15] See the codebooks that document the variable operationalization of the event-based data in Table 3.

[16] See the discussion in the preface of Rummel's book *Death by Government: Genocide and Mass Murder in the Twentieth Century* (1994*a*, xi–xxii) as well as his other books about specific cases. Much of this material is publicly available at Rummel's website: http://www.hawaii.edu/powerkills/welcome.html.

[17] See Poe (2004) for a review of the literature critiquing event-based data. Brysk (1994) provides a specific example critical of comparisons of event-based data.

[18] The seminal work of Lars Schoultz (1981) was the first quantitative test between the stated importance of human rights by the United States government and the allocation of foreign aid using event-based human rights data for Latin American states. The use of event-based counts was criticized (e.g., Carleton and Stohl 1985; Stohl, Carleton and Johnson 1984) and led to a debate about the pros and cons of event-based and standards-based variables. Poe (2004) reviews this debate but interested readers should consult the edited volume by Jabine and Claude (1992) and a symposium on the "Statistical Issues in the Field of Human Rights" published in *Human Rights Quarterly* (Vol. 8, No. 4, 1986).

[19] The models I describe below can be extended to incorporate the actual event counts (Fariss 2013).

[20] Note that both of the models compared by Schnakenberg and Fariss (2014) assume a constant standard of accountability.

[21] All of this research builds on the seminal work by Poole and Rosenthal (1991, 1997), which employs a maximum likelihood

comparison statistics represent the best way to adjudicate between competing theories and the measurement models deduced from them.

To parameterize the changing standard of accountability, I allow the baseline probability of observing a given level of repression for a specific repression variable or *item* (as modeled by time-varying item cut points) to vary as a function of time in one model (dynamic standard model) and compare the resulting estimates to another in which this probability is constant (constant standard model).[22] This is accomplished by estimating time varying "item difficulty cut points" or "thresholds" for some of the items. These parameters are analogous to the intercept term in a linear model. The rest of the model parameters are similar to other latent variable models in the literature and are described in detail below. Thus, the changing standard of accountability is parameterized in the dynamic standard model by estimating the item difficulty cut points for the data sources that are derived from information contained in the annual human rights reports. Constant item difficulty cut points are estimated for the event-based data sources. This parameterization is motivated by the theoretical distinction between the standards-based data sources and event-based data sources. Again, both models are dynamic with respect to the latent variable itself.

Formally, the statistical models I compare in this article are both built on the assumption that the observed repression outcome variables for the country-year observations are each a function of the same underlying unidimensional latent variable, which represents the "true" or "latent" level of repression or respect for physical integrity rights. The goal of these models is to estimate θ_{it}, which is the latent level of respect for physical integrity rights of country i in year t. For each model there are J indicators $j = 1, \ldots, J$. Some of the j indicators are ordinal with varying number of levels and some of the j indicators are binary. As already noted, $i = 1, \ldots, N$ indexes cross-sectional units and $t = 1, \ldots, T$ indexes time periods. y_{itj} is observed for each of the $j = 1, \ldots, J$ indicators displayed in Tables 2 and 3. Each indicator is either ordinal or binary and can take on K_j values. For the binary indicators, $K_j = 2$.

For each item, there is an "item discrimination" parameter β_j and a set of $K_j - 1$ "item difficulty cut points" $(\alpha_{jk})_{k=1}^{K_j}$. These parameters are analogous to a slope and intercept term in a logistic regression or the slope and cut points in an ordered logistic regression.

For the dynamic standard model, I specify the parameterization of the difficulty cut points for some of the items to vary over time such that $(\alpha_{tjk})_{k=1}^{K_j}$. Note the t subscript here. This parameterization includes the standards-based variables from Cingranelli and Richards (1999, 2012*a*, *b*), Gibney, Cornett and Wood (2012), Hathaway (2002), and Conrad, Haglund and Moore (2013). The other items retain the constant item difficulty cut-point parameterization: $(\alpha_{jk})_{k=1}^{K_j}$, which include the binary event-based variables drawn from Harff and Gurr (1988), Harff (2003), Rummel (1994*b*, 1995), Eck and Hultman (2007), Taylor and Jodice (1983).[23] Note the lack of a t subscript here. There is no t subscript on this parameter for any of the items in the constant standard model.

I assume error terms ε_{itj} are independently drawn from a logistic distribution, where $F(\cdot)$ denotes the logistic cumulative distribution function. The probability distribution for a given response to item j in the constant standard model is therefore given by

$$P[y_{itj} = 1] = F(\alpha_{j1} - \theta_{it}\beta_j), \quad \textbf{(1)}$$

$$P[y_{itj} = k] = F(\alpha_{jk} - \theta_{it}\beta_j) - F(\alpha_{jk-1} - \theta_{it}\beta_j), \quad \textbf{(2)}$$

$$P[y_{itj} = K] = 1 - F(\alpha_{jK-1} - \theta_{it}\beta_j). \quad \textbf{(3)}$$

For each item with constant difficulty cut points, $y_{itj} = k$ if $\alpha_{jk-1} < \theta_{it}\beta_j + \varepsilon_{itj} < \alpha_{jk}$, and by specifying $\alpha_{j0} = -\infty$ and $\alpha_{jK_j} = \infty$ the probability equations (1)– (3) reduce to[24]

$$P[y_{itj} = k] = F(\alpha_{jk} - \theta_{it}\beta_j) - F(\alpha_{jk-1} - \theta_{it}\beta_j) \quad \textbf{(4)}$$

Therefore, assuming local independence of responses across units, the constant standard's likelihood function for β, α, and θ, given the data, is $\mathcal{L}(\beta, \alpha, \theta|y)$ and is expressed as

$$\mathcal{L} = \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{j=1}^{J}\left[F(\alpha_{jy_{itj}} - \theta_{it}\beta_j) - F(\alpha_{jy_{itj}-1} - \theta_{it}\beta_j)\right]. \quad \textbf{(5)}$$

The first set of equations (1)– (3) and the reduced form (4) refer to the probability of observing a particular hypothetical level k. The likelihood equation (5) refers to the probability of the observed level in the data y_{itj}. These equations are the same for the dynamic standard model except for the addition of the t subscript on some of the α parameters. As a notational

approach to model political ideology from roll call votes. This model was first used in IR by Voeten (2000) to study the United Nations general assembly. See Clinton, Jackman and Rivers (2004) and Martin and Quinn (2002) for Bayesian implementations of this model. See Jackman (2008) for a thorough discussion of the development of these and other measurement models.

[22] I use the term "item" and "variable" interchangeably throughout this article. The term "item" is attributed to researchers developing educational tests (e.g., Lord 1980; Lord and Novick 1968; Rasch 1980). See Borsboom (2005) and Jackman (2008) for accounts of the development of this literature.

[23] It is a coincidence that the event-based variables are each binary whereas the standards-based data are all categorical. The model is not dependent on this distinction.

[24] For each item with dynamic difficulty cut points, $y_{itj} = k$ if $\alpha_{tjk-1} < \theta_{it}\beta_j + \varepsilon_{itj} < \alpha_{tjk}$, where ε_{itj} is an error term and $\alpha_{tj0} = -\infty$ and $\alpha_{tjk_j} = \infty$.

convenience let $v_j = 1$ when the j indicator is one of the standards-based variables and then $v_j = 0$ when it is one of the event-based variables. The probability distribution for the dynamic standard model is therefore

$$P[y_{itj} = k] = [F(\alpha_{tjk} - \theta_{it}\beta_j) - F(\alpha_{tjk-1} - \theta_{it}\beta_j)]^{(v_j)} * [F(\alpha_{jk} - \theta_{it}\beta_j) - F(\alpha_{jk-1} - \theta_{it}\beta_j)]^{(1-v_j)} \quad \textbf{(6)}$$

and the dynamic standard's likelihood function $\mathcal{L}(\beta, \alpha, \theta|y)$ is expressed as

$$\mathcal{L} = \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{j=1}^{J}\left[F(\alpha_{tjy_{it\,j}} - \theta_{it}\beta_j) - F(\alpha_{tjy_{itj}-1} - \theta_{it}\beta_j)\right]^{(v_j)} * \left[F(\alpha_{jy_{it\,j}} - \theta_{it}\beta_j) - F(\alpha_{jy_{it\,j}-1} - \theta_{it}\beta_j)\right]^{(1-v_j)}. \quad \textbf{(7)}$$

Note that when $v_j = 0$, the probability distribution (6) and the likelihood function (7) for the dynamic standard model are equivalent to equation (4) and (5) for the constant standard model. The model is different when $v_j = 1$, which is when the standard of accountability changes over time.

If θ_{it} was fully observed, then the likelihood functions above would be equivalent to independent ordinal logistic regression models. However, since this is not the case, all of the parameters of interest, the latent variable θ_{it}, the item difficulty cut points α_{tjk} or α_{jk}, and the item discrimination parameters β_j, must be estimated simultaneously so that the model is identified. This issue necessitates the use of Bayesian simulation.

To estimate the models, I set the same priors on the latent variable estimate θ_{it} for both of the models compared in this article such that $\theta_{it} \sim N(\theta_{it-1}, \sigma)$ for all i and t except when $t = 1$, and then is $\theta_{i1} \sim N(0, 1)$. This parameterization captures an old idea in the human rights literature: repression "radiates an after-life" which decreases the need for future repressive actions by the state for a certain period of time.[25] Both of the latent variable models formalize this idea and model comparison statistics help to validate it (Schnakenberg and Fariss 2014).

In both models, I specify $\sigma \sim U(0, 1)$ to reflect prior knowledge that the between-country variation in human rights respect will be higher on average than the average within-country variance.[26] I specify the item-discrimination parameters $\beta_j \sim Gamma(4, 3)$ to reflect the prior belief that all variables contribute significantly and in the same direction to the latent variable. These parameters estimate the strength of the relationship between values of the latent variable and the probability of being coded at a given level for one of the repression variables.[27]

For the dynamic standard model, I relax the assumption about the item difficulty cut points made in other latent variable models and allow the α parameters to vary over time such that the priors of $\alpha_{tjk} \sim N(\alpha_{t-1,jk}, 4)$, subject to the ordering constraint that $\alpha_{tj1} < \alpha_{tj2} < \ldots < \alpha_{tjK-1}$ for all j. When $t = 1$, then $\alpha_{1jk} \sim N(0, 4)$. By allowing the item difficulty cut points for the standards-based variables to vary over time, I am able to assess how the probability of being coded at a specific level on the original ordered repression variables changes from year to year. The priors for the α parameters for the event-based data in the dynamic standard model are $\alpha_{jk} \sim N(0, 4)$, again subject to the same ordering constraint that $\alpha_{j1} < \alpha_{j2} < \ldots < \alpha_{jK-1}$ for all j. This is the same setup for all of the α parameters in the constant standard model. Table 4 summarizes the prior distributions for the model parameters and the differences between their implementation in the dynamic standard model and the constant standard model.[28]

These models, like all item-response theory models, rest on an assumption of local independence. This assumption implies that any two item responses are independent conditional on the latent variable. This means that two item responses are only related because of the fact that they are each an observable outcome of the same latent variable. There are three relevant local independence assumptions: (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. The third assumption is relaxed by incorporating temporal information into prior beliefs about the latent repression variable in both models and the changing standard of accountability in the dynamic standard model, which is captured by the item-difficulty cut points.

Some readers may question the first assumption, which states that different repressive tactics are not causally related to one another within the same country-year but are instead only related to each other through the underlying latent variable. This assumption is made implicitly in other projects that aggregate information about repression into one scale (Cingranelli and Richards 2012*a*, *b*; Gibney, Cornett and Wood 2012; Landman and Larizza 2009). Jackman (2008) and van Schuur (2003) provide further details about this assumption. More importantly however, different repressive tactics can be related to one another in theoretically important but noncausal ways. Fariss

[25] See the quote by Duvall and Stohl (1983), which is cited by Stohl et al. (1986).

[26] This is not a consequential decision in terms of restricting the values of this parameter because the posterior estimates of σ from the converged model is less than 0.05, making the truncation decision unimportant.

[27] Prior sensitivity analyses suggested that this was not restrictive. When normal priors were specified for each β, the posterior densities rarely overlapped with zero. However, a model without this restriction is not identified with respect to rotation.

[28] Each model compared in this article is estimated with two MCMC chains, which are run 100,000 iterations using JAGS (Plummer 2010) on the Gordon Supercomputer (Sinkovits et al. 2011). The first 50,000 iterations were thrown away as burn-in and the rest were used for inference. Diagnostics all suggest convergence (Geweke 1992; Heidelberger and Welch 1981, 1983; Gelman and Rubin 1992).

TABLE 4. Summary of Prior Distributions for Latent Variable and Model Level Parameter Estimates

Parameters	Constant Standard	Dynamic Standard
Country-year latent variable (first year)	$\theta_{i1} \sim N(0, 1)$	$\theta_{i1} \sim N(0, 1)$
Country-year latent variable (other years)	$\theta_{it} \sim N(\theta_{it-1}, \sigma)$	$\theta_{it} \sim N(\theta_{it-1}, \sigma)$
Uncertainty of latent variable	$\sigma \sim U(0, 1)$	$\sigma \sim U(0, 1)$
Event-based variable cut points (constant)	$\alpha_{jk} \sim N(0, 4)$	$\alpha_{jk} \sim N(0, 4)$
Standards-based variable cut points (constant)	$\alpha_{jk} \sim N(0, 4)$	———————
Standards-based variable cut points (first year)	———————	$\alpha_{1jk} \sim N(0, 4)$
Standards-based variable cut points (other years)	———————	$\alpha_{tjk} \sim N(\alpha_{t-1,jk}, 4)$
Slope	$\beta_j \sim Gamma(4, 3)$	$\beta_j \sim Gamma(4, 3)$

Note: Both of these models are dynamic in the treatment of the latent variable θ.

and Schnakenberg (2013), following many analysts before them,[29] assume that repression is a useful tool for a leader because it produces the benefit of mitigating potential threats to the regime. However, the emphasis of this theory is that many repressive behaviors may be complementarity policy options. A "complement" is defined if the presence of one repressive policy tool reduces the probability that the use of another policy tool is made public, decreases the threat the first policy tool was used to address, or reduces the possibility of retribution faced by a leader caught using the original policy tool.[30] This theoretical distinction emphasizes the choices of the policy maker in selecting repressive tools. It is the leader selecting to both torture and imprison political opponents because of a threat, which is the underlying cause of repression generally and the two repressive behaviors specifically. This is an important theoretical and empirical issue that human rights scholars are currently grappling with (e.g., Conrad and Demeritt 2011; Fariss and Schnakenberg 2013).

Recent human rights scholarship has begun to analyze both the interrelationships between the different state repressive behaviors captured by the CIRI data (e.g., Conrad and Demeritt 2011; Fariss and Schnakenberg 2013). However, these scholars are quick to point out that no definitive answer has been reached about the complimentarily vs. substitutability of these state behaviors. Fariss and Schnakenberg (2013) only analyze the systemwide level of co-occurance between different CIRI rights. Conrad and Demeritt (2011) focus on the 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" (2011, 14). The evidence presented by Conrad and Demeritt (2011) that state leaders choose to substitute one type of abuse for the other is consistent with the assumption of the models in that the relationship is not directly causal but instead dependent on the underlying latent trait, which is affected directly by the strategy of the political authorities.

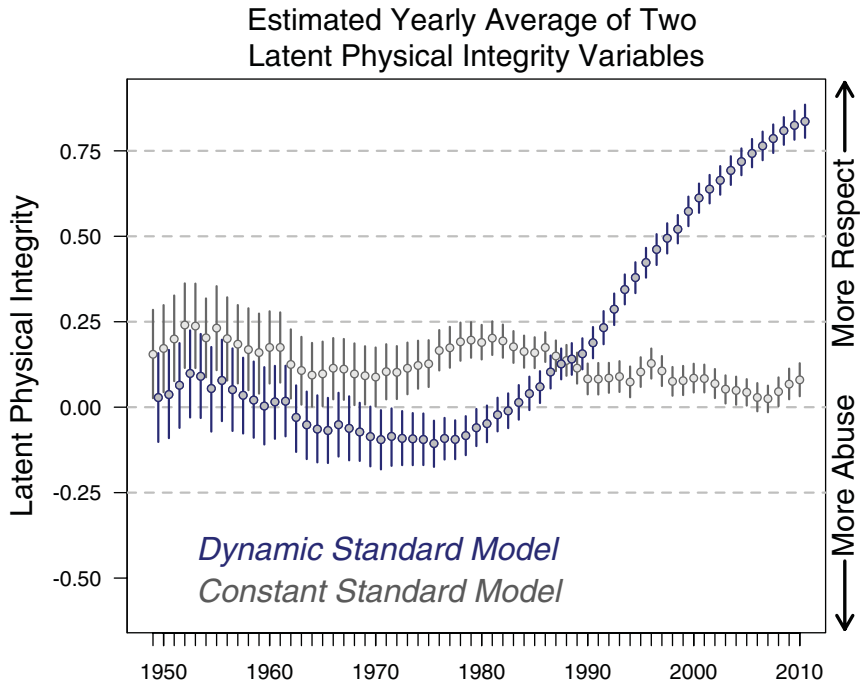
The model developed in this article can be extended to help address this issue. Such an extension with its requisite theoretical justification is outside the scope of this article. However, I have estimated alternative versions of the dynamic standard model and the constant standard model which relax the assumption of local independence of different indicators within the same country-year. The latent variables estimates produced by these models follow the same pattern as the latent variable estimates presented below. Briefly in the appendices I describe this model and several alternatives, which might be useful extensions for future research (see Appendix P). Overall, these alterations do not change the main inference of the article, that the standard of accountability has become more stringent over time and that the average level of human rights has improved once this confounding factor is taken into account.

The models developed above assume that repression is caused by choices made by the regime that lead to some "true" level of repression in each country, each year. This is the latent variable of repression, which the models attempt to estimate based on observable outcomes. The observables are each a function of the latent variable, which are captured in the content of documents produced by human rights analysts working for Amnesty International and the U.S. State Department. This content is then coded into data by research analysts (i.e., Cingranelli and Richards 1999, 2012*a*, *b*; Conrad, Haglund and Moore 2013; Gibney, Cornett and Wood 2012; Hathaway 2002). If the standard of accountability that the monitors use when assessing state behaviors changes over time, then data derived from these documents will be biased. Though the standard of accountability is not directly observable, I have parameterized it in the dynamic standard model, which can be interpreted as the baseline probability of observing a

[29] See, for example, Carey (2006); Davenport (2007*b*); Mason and Krane (1989); Moore (1998, 2000); Poe and Tate (1994); Poe, Tate and Keith (1999); Poe (2004); Zanger (2000).

[30] Fariss and Schnakenberg (2013) find that, on average, physical integrity rights abuses are complements with one another each year using the country-year CIRI variables (1981–2006).

FIGURE 3. Yearly Mean and Credible Intervals for Latent Physical Integrity Estimates from Two Models



Notes: The dynamic standard model allows the $k - 1$ difficulty cut points, the baseline probability of being coded at a certain level on the original standards-based repression variables, to vary over time. The standards-based variables are those which use human rights reports from the U.S. State Department or Amnesty International as their primary information source. The model with a constant standard estimates one set of $k - 1$ cut points for every repression variable including the standards-based variables. The difference in the two sets of estimates suggests that an increasing standard of accountability explains why the average level of repression has remained unchanged over time when the changing standard is not taken into account. By allowing this standard to vary with time, a new picture emerges of improving physical integrity practices over time, which begins after initially deteriorating from the beginning of the period until the late 1970s. Appendix E contains selected country examples similar to this figure.

given level of repression for a specific standards-based variable (as modeled by time-varying item cut points). If, for example, the probability of reporting frequent levels of torture increases from year to year, then this is evidence of the changing standard of accountability as captured by these documents. In the next section, I compare the model estimates from this new model, the dynamic standard model, to those from the constant standard model, which allows me to demonstrate that the standard of accountability has increased over time for several of the human rights measures derived from the content of the human rights reports from Amnesty International and the U.S. State Department.

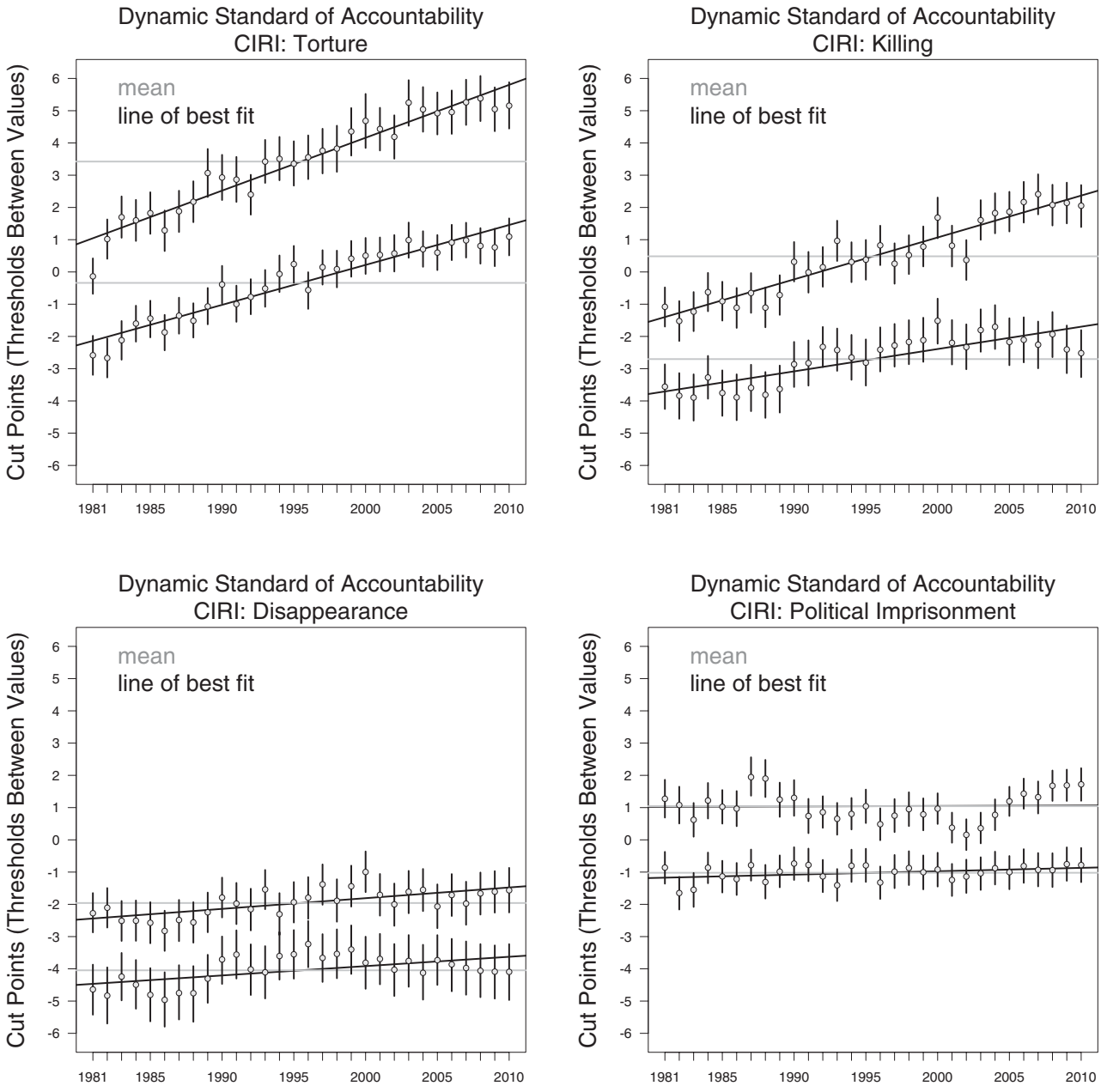
RESULTS: PHYSICAL INTEGRITY PRACTICES HAVE IMPROVED

A comparison of latent variable estimates from the dynamic standard model with those from the constant standard model provide strong evidence for a new view of repression: *physical integrity practices have improved over time*. Unobserved changes to the standard of accountability explain why average levels of repression have appeared to remain unchanged as the constant standard model would suggest. Differences in the average level of the latent variable estimates are displayed in Figure 3. For the constant standard model to be more consistent with reality and for this same pattern to obtain, the monitoring agencies would need to produce the human rights reports consistently from year to year *and* the producers of the event-based data would need to use a less and less stringent definition of repression in the assessment of these events over time. Neither of these alternative behaviors are supported by the theory nor the model comparison tests, which I introduce below. First, I discuss how the standard of accountability manifests for different standards-based variables.

How does the changing standard of accountability influence the probability of being coded at a specific level of repression for the original standards-based variables over time? Figures 4 and 5 present panels that each display changes in the item difficulty cut points (thresholds between values for each repression variable) from the dynamic standard model for each of the eight variables derived from the standards-based reports.

The changing standard of accountability does not affect all of the standards-based variables equally. Countries are far more likely to be coded for frequent torture based on the CIRI coding rules today than countries

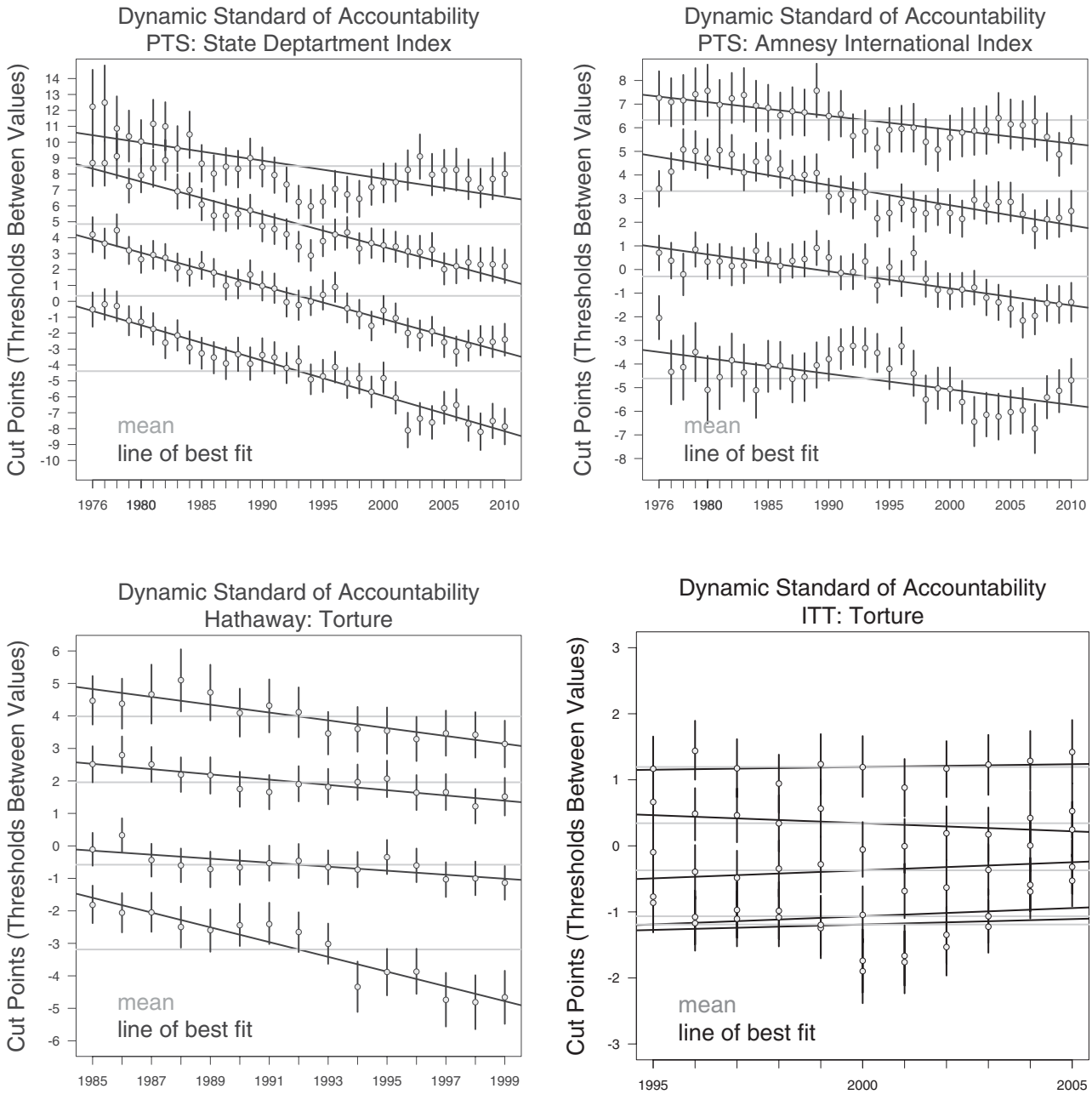
FIGURE 4. Visualizing the Changing Standard of Accountability



Notes: An increase in the difficulty cut points translates directly into a statistically significant change in the probability of being classified as a 0, 1, or 2 on the original CIRI variables such that being classified as 0 (e.g., frequent abuse) becomes more likely and 2 (e.g., no abuse) becomes less likely as a function of time. There is no statistical relationship between the Political Prison variable and time. See Appendix G for the posterior estimates of these parameters and Appendix F for additional figures.

with similar levels of repression just a few decades ago. As the standard of accountability becomes more stringent, monitoring agencies look harder for torture, look in more places for torture, and classify more acts as torture. All of the standards-based variables, with the exception of the CIRI imprisonment variable and the ITT torture variable, are affected by changes to the standard of accountability (see Appendices F and G for more details). However, as Clark (2001) discusses in her book, the original mission of Amnesty International was to document political imprisonment. The documentation of other human rights abuses came about as states responded to the advocacy efforts of Amnesty and other human rights NGOs. It is not surprising that the human rights reports consistently document political imprisonment over time. The lack of temporal change in the probability of coding levels of torture in the ITT data may reflect the relatively short

FIGURE 5. Visualizing the Changing Standard of Accountability



Notes: A decrease in the difficulty cut points translates directly into a statistically significant change in the probability of being classified as a 1, 2, 3, 4, or 5 on the original political terror scale variables and the hathaway torture variable such that being classified as 1 (e.g., little to no abuse) becomes less likely and 5 (e.g., widespread abuse) becomes more likely as a function of time. There is no statistical relationship between the ITT variable and time. See Appendix G for the posterior estimates of these parameters and Appendix F for additional figures.

period of coverage (1995–2005) or differences between Amnesty's Urgent Action Reports, upon which these data are based, and the annual report used by the other data sources. Additional analysis is necessary on this specific issue.

The lack of results for these two variables is actually quite encouraging for the plausibility of the dynamic standard model. In effect, these two variables, in addition to the five event-based indicators, acted as a baseline for the model so that both the overall level of repression and the changing standard of accountability could be estimated simultaneously. These results help to alleviate concern that the changing standard of accountability is an unwanted artifact rather than a theoretically specified feature of the model.

MODEL COMPARISONS

Deviance Information Criteria (DIC)

Next, I test to see if the dynamic standard model is a better approximation of reality relative to the alternative constant standard model. Readers should keep in mind that all existing models of repression—those based on information from the annual human rights reports—make the same assumption about a constant standard of accountability over time. Models of repression include all of the existing human rights scales that aggregate information about different rights abuses from the annual human rights reports (Cingranelli and Richards 1999, 2012*a*, *b*; Gibney, Cornett and Wood 2012; Hathaway 2002) in addition to the latent variable model recently developed by Schnakenberg and Fariss (2014) and the factor analytic method used by Landman and Larizza (2009). By comparing the estimates from the two latent variable models, I am able to test the hypothesis that an increase in the standard of accountability—the probability of observing and therefore coding a repressive outcome (as modeled by time-varying item cut points)—increases over time for the repression variables derived from the human rights reports.

For these tests, I first present a statistic called the Deviance Information Criterion or DIC for short. This statistic provides information analogous to a penalized likelihood ratio test or simply the comparison of adjusted-R^2 statistics from competing models. For the DIC statistic, relatively smaller values indicate that a model explains more of the variance in outcome variables compared to an alternative model. Recall that the outcome variables in both models are the original repression variables. The DIC statistic is a method useful for comparing the models in this article because it penalizes more complex models so that the more parsimonious one is favored, all else equal (Gelman et al. 2003). Thus a smaller DIC for this more complex model is strong evidence of its improvement over alternatives. Spiegelhalter, Best, Carlin and Van Der Linde (2002) proposed that differences of greater than 5 or 10 provide substantial evidence in favor of the model with the lower DIC. The DIC statistics are 53,706 for the dynamic standard model and 55,027 for the constant standard model, which is a difference of several thousand in favor of the dynamic standard model. See Appendix I for more details.

Posterior Predictive Checks

Posterior predictive checks assess the quality of the model by direct comparison of model predictions from competing models. These tests compare predictions of the original repression variables generated by the two competing latent variable models. The results suggest that the dynamic standard model again outperforms its competitor.

At every iteration of the MCMC algorithm, the model parameters can be used to make a prediction of each of the observed repression variables included in the model. The better fitting model should on average generate predictions closer to the observed values of these variables when compared to similar predictions from a competing model (Gelman and Hill 2007).

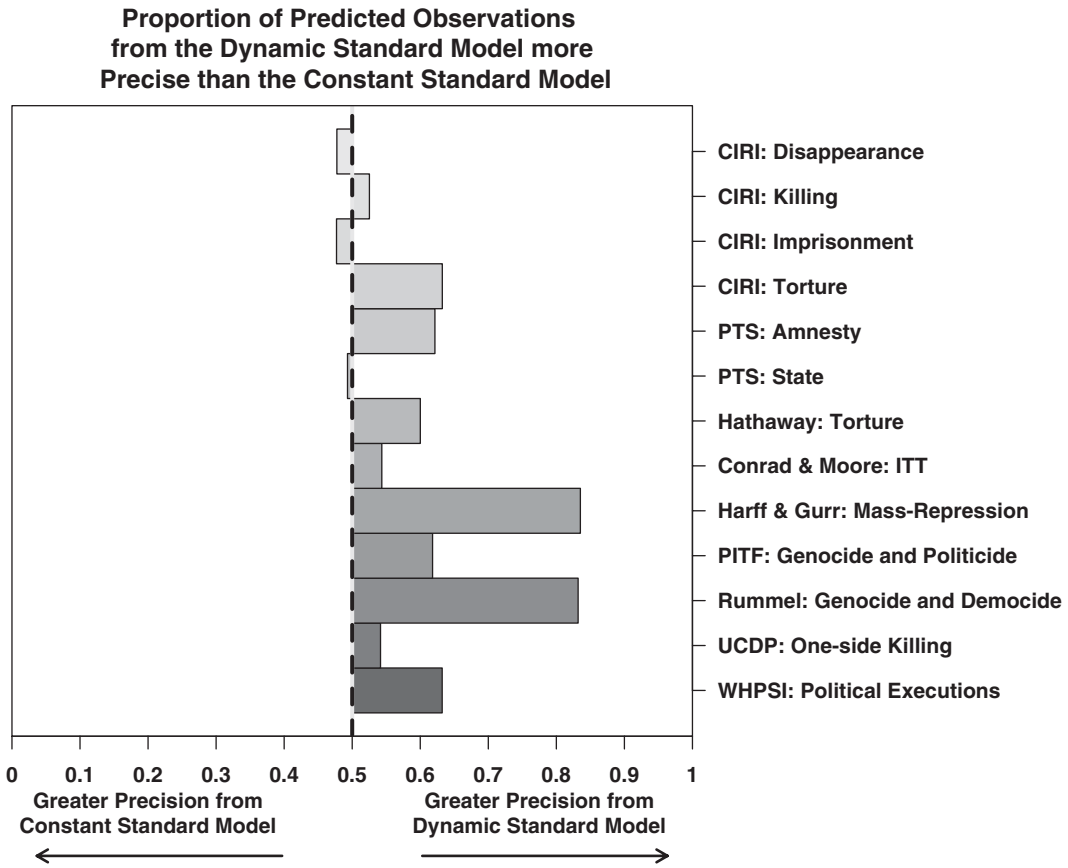
Formally, for each draw from the posterior distribution, I predict each of the j items $y_{it\,j}$ for every country-year observation for which $y_{it\,j}$ is observed. Since there are thousands of draws from the posterior distribution, indexed by d, I am able to calculate the sum of squared differences of observed $y_{it\,j}$ and the posterior predicted values $\hat{y}^{dynamic}_{it\,j\,d}$ from the dynamic standard model using the equation: $S^{dynamic}_{it\,j} = \sum_d (y_{it\,j} - \hat{y}^{dynamic}_{it\,j\,d})^2$ and likewise for the posterior predicted values $\hat{y}^{constant}_{it\,j\,d}$ from the constant standard model: $S^{constant}_{it\,j} = \sum_d (y_{it\,j} - \hat{y}^{constant}_{it\,j\,d})^2$

I have aggregated the sum of squared difference for each observation to compare values for the same observation from the competing models. These comparisons are captured in Figure 6 and a table in Appendix J. Figure 6 displays the proportion of observations such that : $S^{dynamic}{}_{it\,j} \leq S^{constant}{}_{it\,j}$, or in words, when the sum of squared difference from the dynamic standard model is $\leq$ to the constant standard model for each country-year observation for all of the repression variables. Proportions closer to 1 indicate that the dynamic standard model outperforms the constant standard model at predicting the original repression variables. Proportions closer to 0 indicate the opposite. Proportions at 0.50 indicate that both models are predicting the items with about the same amount of error relative to each other. The proportions increase as the number of observations with a smaller sum of squared deviation increases when comparing the dynamic standard model and the constant standard model. The dynamic standard model does a much better job of predicting the original repression variables, especially the event-based variables. The improvement occurs for the event-based variables because of the temporal bias that exists in the standards-based variables. The constant standard model does not account for this bias, which reduces its ability to accurately predict the values of the event-based data not affected by the changing standard of accountability.

THE CHANGING STANDARD OF ACCOUNTABILITY AND TREATY COMPLIANCE: THE CASE OF THE CONVENTION AGAINST TORTURE

In this section, I illustrate the substantive importance of the changing standard of accountability for international relations theory by showing that ratification of the UN Convention Against Torture and respect for physical integrity rights is positive. This result contradicts negative findings from existing research. As the standard of accountability has increased over time, empirical associations with human rights data derived from standards-based documents and other variables will be biased if changes in the human rights documents are not accounted for. This is

FIGURE 6. Proportion of Predicted Observations



Notes: Proportions closer to 1 indicate that the dynamic standard model outperforms the constant standard model at predicting the original repression variables. Proportions closer to 0 indicate the opposite. Proportions at 0.50 indicate that both models are predicting the items with about the same amount of error relative to each other. The dynamic standard model does a much better job of predicting the original repression variables, especially the event-based variables.

especially true for variables that measure the existence of institutions that are correlated with time, such as whether or not the UN Convention Against Torture has been ratified.

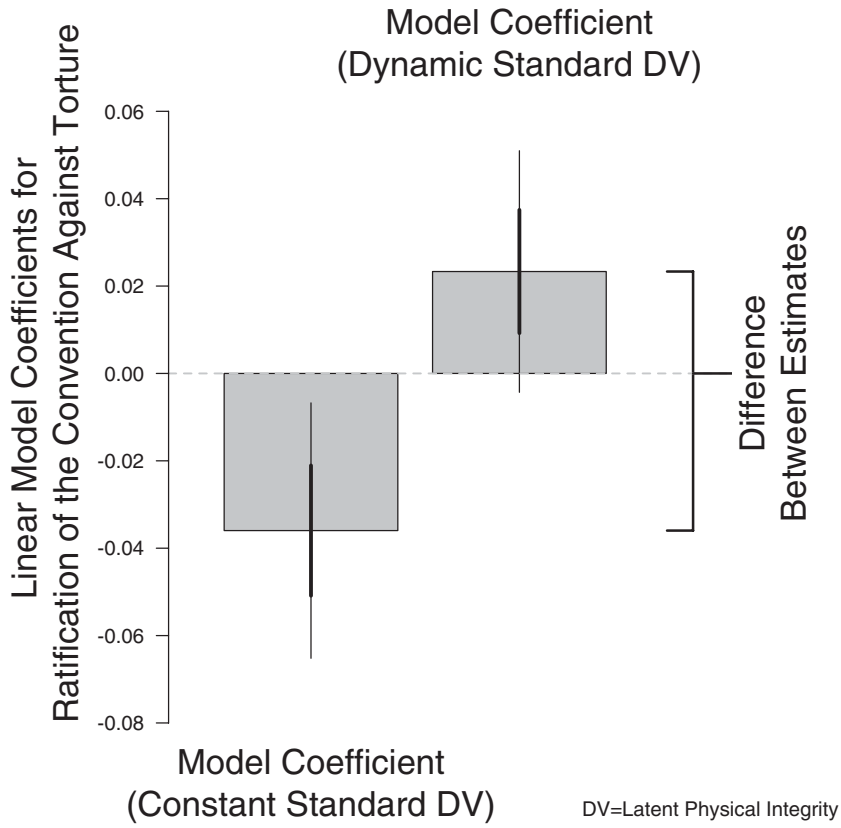
In the international relations literature there are two opposing viewpoints on treaty effectiveness. Authors such as Morrow (2007), Simmons (2000), and Simmons and Hopkins (2005) argue that treaty ratification constrains actors to modify their behaviors by creating costs for noncompliance. An alternative viewpoint is that countries only ratify a treaty if they would have complied even in the absence of the treaty. Thus, treaties have no effect on the behavior codified within the treaty, such as the level of cooperation (e.g., Downs, Rocke and Barsoom 1996; Von Stein 2005), or ratification of certain human rights treaties (e.g., Hafner-Burton and Ron 2009; Hafner-Burton and Tsutsui 2005, 2007; Hathaway 2002). The results presented here call into question this second viewpoint. The new latent variable model I have developed provides a way to improve the measurement of respect for human rights specifically and potentially the measurement of other forms of compliance in international relations more generally.

For this demonstration, I compare linear model coefficients using the latent variable from the constant standard model and the latent variable from the dynamic standard model. That is, I estimate two linear regression equations using the latent physical integrity variables from the two measurement models as dependent variables. I regress these variables on a binary variable that measures whether or not a country has ratified the Convention Against Torture in a given year. I also include several control variables.[31]

New inferences are obtained by simply replacing the dependent variable derived from the constant standard model with the one from the dynamic standard model. Figure 7 plots the coefficient for CAT ratification from the linear models, which each use one of the two latent physical integrity dependent variables. The linear regression using the dependent variable from the constant standard model generates a negative coefficient,

[31] I include measures of democracy (Marshall, Jaggers and Gurr 2013), the natural log of GDP per capita (Gleditsch 2002), the natural log of population, and the lagged value of the latent variable. See Appendix N for more information about this and other specifications in addition to information about incorporating uncertainty inherent in the lagged latent variable.

FIGURE 7. Estimated Coefficient for CAT (the UN Convention Against Torture) Ratification From the Linear Model using the Dependent Latent Physical Integrity Variables From the Constant Standard Model and the Dynamic Standard Model, Respectively.



Notes: The thick lines represent 1 ± the standard error of the coefficient. The thin lines represent 2 ± the standard error of the coefficient. The difference between the coefficients is statistically significant ($p < 0.004$).

which corroborates results from earlier work. Comparison with the regression coefficient from the model using the dependent variable from the dynamic standard model is striking. The coefficient has flipped signs and is statistically significant when compared with 0 ($p < 0.098$) and the alternative coefficient ($p < 0.004$). These results suggest that human rights protectors are more likely to ratify the treaty, that the treaty may in fact have some causal effect on human rights protection, or possibly both. Overall, these findings suggest that the treaty is not merely cover for human rights abusers.

Note that these models are not designed for causal inference and, though a variety of selection issues are known to exist when using this specification,[32] the results from this type of model have spawned a large literature because of the counterintuitive, negative correlation found between ratification and respect for human rights (e.g., Hafner-Burton and Tsutsui 2005, 2007; Hathaway 2002; Hollyer and Rosendorff 2011; Vreeland 2008). Though this finding has been criticized (Clark and Sikkink 2013; Goodman and Jinks 2003), it is generally taken for granted in the literature (Hafner-Burton and Ron 2009). Importantly, this new result calls into question a key assumption about state behavior made in several recent articles (e.g., Hollyer and Rosendorff 2011; Vreeland 2008) and a book (Hafner-Burton 2013) on human rights treaty compliance. The conclusions and policy recommendations from this work should be re-evaluated.

Overall, much additional testing is needed to probe the differences between existing empirical relationships and the new relationships generated using the latent physical integrity estimates derived from the dynamic standard model. A recent manuscript uses cross-validation and random forest methods to determine the predictive power of the covariates identified as important in the literature on repression using the existing CIRI and PTS physical integrity scales and the

[32] See discussions in Neumayer (2005), Simmons and Hopkins (2005), Von Stein (2005), Simmons (2009), Hill (2010), and most recently Conrad and Ritter (2013) and Lupu (2013*b*). The selection issue that these authors address is orthogonal to the differences in the two latent variable models. It is therefore sufficient to use this illustration to demonstrate how different inferences are obtained using the latent variable from the dynamic standard model.

new estimates presented in this article (Hill and Jones 2014). The cross-validation and random forest methods corroborate the result that ratification of the Convention Against Torture is positively associated with the new latent variable generated from the dynamic standard model presented in this article. The authors also find that measures of civil war (e.g., Davenport 2007*a*; Poe and Tate 1994), "youth bulges" (Nordås and Davenport 2013), domestic legal institutions (Keith, Tate and Poe 2009), and state reliance on natural resource rents (Demeritt and Young 2013), are good predictors of levels of repression. I also examine the relationship between the new latent variable estimates and other human rights treaties in another article (Fariss 2014). The new model introduced in this article might also be useful for analyzing other issue areas of treaty compliance in international relations, which I leave for future research.

SUGGESTIONS FOR FUTURE RESEARCH

If changes in the standard of accountability are not addressed in applied research, then biased inferences are the likely result. Bias occurs because of the increasing number of years for which standards-based variables exist. Figure 8 captures the increasing disagreement between the latent variables estimates generated from the dynamic standard model and those from the constant standard model (1976–2010). The disagreement occurs because the dynamic standard model incorporates the changing standard of accountability, whereas the constant standard model, which is biased, does not.

The first option for analysts is to simply use the new latent repression estimates from the dynamic standard model. As I demonstrated in Section 7, a linear model can easily accommodate the latent repression estimates as the dependent variable. Schnakenberg and Fariss (2014) describe a method for incorporating the uncertainty associated with the latent variable estimates in this model or any other model that uses the lagged latent variable estimates as an independent variable (see Appendix L for more details).

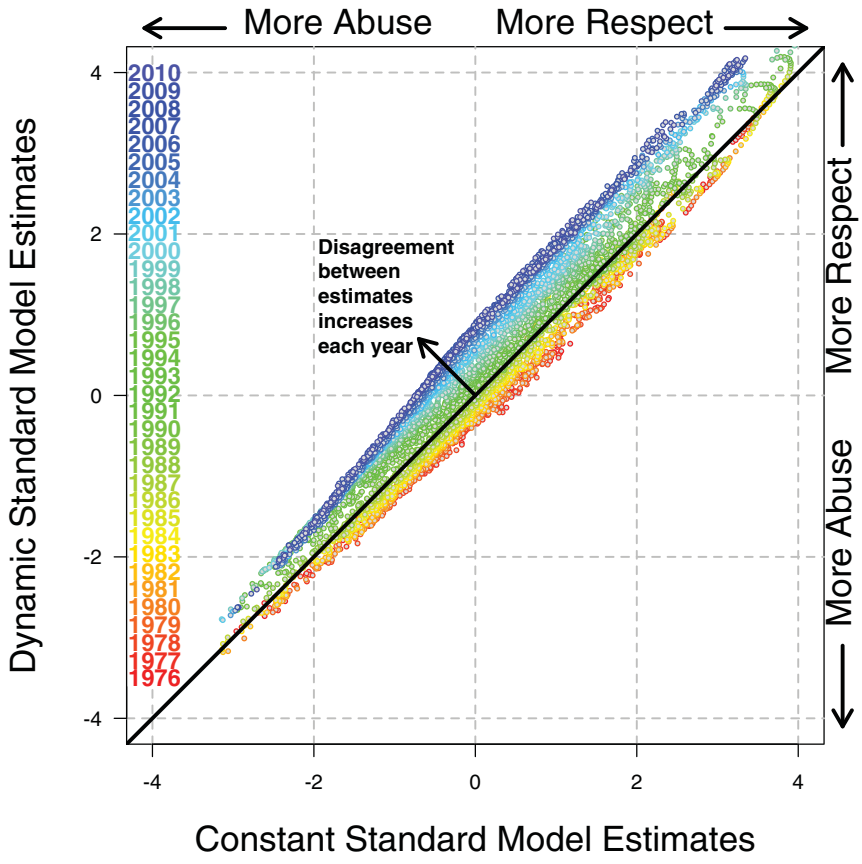
Analysts interested in any of the standards-based variables as a dependent variable should consider using a hierarchical model with the lagged estimate of repression generated from the dynamic standard model in addition to specifying time-varying cut points. This specification will help to avoid generating biased inferences. Through Bayesian simulations, programs such as JAGS, Stan, or WinBUGS can handle this more difficult to estimate model when using the standards-based variables. The alternative to this approach still involves specifying a time variable (a count of the number of years in the study beginning with the first year) interacted with the lagged repression estimates generated in this article. In the appendix (Appendices L and M), I describe the specification for models using the original standards-based variables. I also present a procedure for modeling the original binary event data. These analyses also generate additional predictive validity statistics that corroborate the results from the DIC statistics and posterior predictive checks presented above.

CONCLUSION

By allowing the standard of accountability to vary with time, a new picture emerges of improving physical integrity practices over time. Recall the research question I posed at the beginning of this article: *Have levels of political repression changed?* To answer this question, I argued that the use of repressive policy tools appears unchanged over time because of the unaccounted-for standard of accountability that monitoring agencies use to hold states responsible for abuse. I theorized that the standard of accountability, which I defined as a set of expectations or norms that state behaviors are measured against, changed over time because of the combination of three tactics that make up the strategies of monitoring agencies. The standard of accountability changes over time because of the incentives of reporting agencies to (1) gather more accurate information about credible allegations of repression, (2) broaden the coverage of information gathering campaigns with the help of other NGOs, and (3) continually press governments to reform through naming and shaming campaigns, even after real reforms are implemented to reduce more egregious rights violations by those governments. The theory allowed me to parameterize a measurement model of repression that incorporates the changing standard of accountability by allowing the baseline probability of observing a given level of repression (as modeled by time-varying item cut points) for a specific repression variable to vary over time.

The results provide strong evidence that the changing standard of accountability affects the content of the human rights country reports produced annually by Amnesty International and the U.S. State Department. The answer to the question posed above is that respect for physical integrity rights has improved after all. Put another way, the level of repression has decreased in the system over time, but this change was masked in the text of the human rights documents by a confounding factor—the standard or accountability—for which researchers had not previously accounted. By accounting for this additional factor, a new picture of global repression emerges in which conditions actually improve over the period of study (1949–2010), since hitting a low point in the mid-1970s. This result has implications for the research agenda of many human rights scholars and should not be left unaddressed in future research. In Section 7, I demonstrated that the empirical relationship between the new physical integrity variable and ratification of the UN Convention Against Torture is positive, which contradicts results from earlier research. Any variable that changes over time and leads to a difference in the level of protection for human rights, like ratification of the Convention Against Torture, is working against temporal bias in the existing human rights scales. Thus, the new estimates I have generated as part of this project can be used to reassess other empirical relationships in the

FIGURE 8. Relationship Between the Latent Variable Estimates Generated From the Dynamic Standard Model on the *y* Axis and the Estimates Generated From the Constant Standard Model on the *x* Axis (1976–2010)



Notes: The 45-degree line represents perfect agreement between the two estimates. Disagreement between the two sets of estimates increases as a function of time.

quantitative human rights literature. In addition to using the new latent variable estimates, it is also possible to reassess relationships with the existing scales. Briefly in Section 8 and in more detail in the appendices, I discussed several methods for addressing the issue of temporal bias in empirical research that uses existing standards-based data and event-based data as dependent variables. The new latent variable model I have developed provides a way to improve the measurement of respect for human rights specifically and potentially the measurement of other forms of treaty compliance or other behaviors in international relations more generally.

To close, I wish to emphasize that the theory and model developed in this article are not meant as a critique of any of the standards-based variables themselves. As should be clear, the theory and the model derived from it are focused solely on the changing standard of accountability, which influences the strategies used by monitoring agencies to generate primary source documents. It is these monitoring agencies and the documents they produce which are under investigation. In fact, this article is itself a testament to the quality of the standards-based data because each of the variables included in the analysis reliably and accurately operationalizes content from these reports.[33]

However, as Clark and Sikkink (2013) discuss, the coding scheme itself may not accurately capture all of the variation in human rights levels that exist in reality. However, I believe that the coding of the reports is reliable. By reliable, I mean that the CIRI, PTS, and Hathaway variables consistently represent the content of the human rights reports published annually by the U.S. State Department and Amnesty International, conditional on the scheme itself. The argument made by Clark and Sikkink is about the validity of the CIRI and PTS variables relative to the theoretical construct of interest, which is respect for human rights. If the PTS or CIRI teams were only interested in accurately measuring the content of the reports then there would be little reason for Clark and Sikkink to question the validity of the resulting variables. The validity comes

[33] Each of the data sources report reliability statistics in their respective code books.

into questions when researchers make the conceptual leap from variables which are based on coded reports to the assumption that the values of those variables represent the "true" level of human rights abuse. This is an important theoretical distinction which is often overlooked when the PTS, CIRI, and Hathaway variables are presented as measurements of *abuse* instead of *reported abuse*.

What this theoretical distinction means in practice however, is that the coded values from the human rights reports come to mean something different when the reports change systematically. This is especially the case for over time changes, which the dynamic standard model addresses. My analysis, though not about content analysis per se, is focused on the comparison of coded human rights documents that have systematically changed over time. However, as the tools of automated content analysis become more popular and accepted in political science, it will be all the more important to determine if the documents selected for analysis are comparable. That is, have the documents systematically changed thus biasing the resulting codings? If so, what are the solutions to such issues? In this article, I have offered an applied solution for this problem in the context of hand-coded human rights documents, which in principle could be adapted to other corpuses of expert-coded or machine-coded text. One promising area of research in which the dynamic standard model could prove of direct use is in the measurement of democracy. Though much has been written about the measurement of democracy (e.g., Gleditsch and Ward 1997; Przeworski, Alvarez, Cheibub and Limongi 2000), there are few systemic assessments of how the conceptualization and therefore the values of the Polity scale or alternative latent democracy variables (e.g., Pemstein, Meserve and Melton 2010; Treier and Jackman 2008) change over time.

REFERENCES

Amnesty International. 2006. *Amnesty International's Country Dossiers and Publications, 1962–2005*. Leiden: IDC Publishers. http://www.idcpublishers.com/ead/ead.php?faid=127faid.xml

Apodaca, Clair. 2001. "Global Economic Patterns and Personal Integrity Rights after the Cold War." *International Studies Quarterly* 45 (4): 587–602.

Armstrong, David A. 2011. "Stability and Change in the Freedom House Political Rights and Civil Liberties Measures." *Journal of Peace Research* 48 (5): 653–62.

Bell, Sam, K. Chad Clay, and Amanda Murdie. 2012. "Neighborhood Watch: Spatial Effects of Human Rights INGOs." *Journal of Politics* 74 (2): 354–68.

Berman, Maureen R., and Roger S. Clark. 1982. "State Terrorism: Disappearances." *Rutgers Law Journal* 13 (3): 531–78.

Bollen, Kenneth A. 1986. "Political Rights and Political Liberties in Nations: An Evaluation of Human Rights Measures, 1950 to 1984." *Human Rights Quarterly* 8 (4): 567–91.

Borsboom, Denny. 2005. *Measuring the Mind*. Cambridge, England: Cambridge University Press.

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

Bueno De Mesquita, Bruce, Feryal M. Cherif, George W. Downs, and Alastair Smith. 2005. "Thinking Inside the Box: A Closer Look at Democracy and Human Rights." *International Studies Quarterly* 49 (3): 439–58.

Carey, Sabine C. 2006. "The Dynamic Relationship Between Protest and Repression." *Political Research Quarterly* 59 (1): 1–11.

Carleton, David, and Michael Stohl. 1985. "The Foreign Policy of Human Rights: Rhetoric and Reality from Jimmy Carter to Ronald Reagan." *Human Rights Quarterly* 7 (2): 205–29.

Cingranelli, David L., and Mikhail Filippov. 2010. "Electoral Rules and Incentives to Protect Human Rights." *Journal of Politics* 72 (1): 243–57.

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.

Cingranelli, David L., and David L. Richards. 2012*a*. "The Cingranelli-Richards (CIRI) Human Rights Data Project Coding Manual Version 2008.3.13." http://ciri.binghamton.edu/documentation/ciri_coding_guide.pdf

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

Clark, Ann Marie. 2001. *Diplomacy of Conscience*. Princeton, NJ: Princeton University Press.

Clark, Ann Marie, and Kathryn Sikkink. 2013. "Information Effects and Human Rights Data: Is the Good News about Increased Human Rights Information Bad News for Human Rights Measures?" *Human Rights Quarterly* 35 (3): 539–68.

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?" Presented at the Annual Meeting of the American Political Science Association.

Conrad, Courtenay R., Jillienne Haglund, and Will H. Moore. 2013. "Disaggregating Torture Allegations: Introducing the Ill-Treatment and Torture (ITT) Country-Year Data." *International Studies Perspectives* 14 (2): 199–220.

Conrad, Courtenay R., and Will H. Moore. 2010. "What Stops the Torture?" *American Journal of Political Science* 54 (2): 459–76.

Conrad, Courtenay R., and Will H. Moore. 2011. "The Ill-Treatment & Torture (ITT) Data Project (Beta) Country-Year Data User's Guide." *Ill Treatment and Torture Data Project*. http://www.politicalscience.uncc.edu/cconra16/UNCC/Under_the_Hood.html

Conrad, Courtenay R., and Emily Hencken Ritter. 2013. "Treaties, Tenure, and Torture: The Conflicting Domestic Effects of International Law." *Journal of Politics* 75 (2): 397–409.

Dancy, Geoff, and Kathryn Sikkink. 2012. "Ratification and Human Rights Prosecutions: Toward a Transnational Theory of Treaty Compliance." *NYU Journal of International Law and Politics* 44 (3): 751–90.

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.

Davenport, Christian. 2007*a*. "State Repression and Political Order." *Annual Review of Political Science* 10: 1–23.

Davenport, Christian. 2007*b*. "State Repression and the Tyrannical Peace." *Journal of Peace Research* 44 (4): 485–504.

Davenport, Christian. 2010. *State Repression and the Domestic Democratic Peace*. New York: Cambridge University Press.

Davenport, Christian, and David A. Armstrong. 2004. "Democracy and the Violation of Human Rights: A Statistical Analysis from 1976 to 1996." *American Journal of Political Science* 48 (3): 538–54.

Demeritt, Jacqueline H.R. and Joseph K. Young. 2013. "A Political Economy of Human Rights: Oil, Natural Gas, and State Incentives to Repress." *Conflict Management and Peace Science* 30 (2): 99–120.

Downs, George W., David M. Rocke, and Peter N. Barsoom. 1996. "Is the Good News About Compliance Good News About Cooperation?" *International Organization* 50 (3): 379–406.

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

Eck, Kristine, and Lisa Hultman. 2007. "Violence Against Civilians in War." *Journal of Peace Research* 44 (2): 233–46.

Fariss, Christopher J. 2013. "Uncertain Events: A Dynamic Latent Variable Model of Human Rights Respect and Government

Killing with Binary, Ordered, and Count Outcomes." Presented at the Annual Meeting of the Society for Political Methodology.
Fariss, Christopher J. 2014. "Human Rights Treaty Compliance and the Changing Standard of Accountability." *Working Paper*.
Fariss, Christopher J., and Keith Schnakenberg. 2013. "Measuring Mutual Dependence Between State Repressive Actions." *Journal of Conflict Resolution*. DOI:10.1177/0022002713487314.
Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 2003. *Bayesian Data Analysis, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC: London.
Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press.
Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science* 7: 457–511.
Geweke, John. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments." In *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and J. F. M. Smith. Oxford: Oxford University Press, 169–93.
Gibney, Mark, Linda Cornett, and Reed M. Wood. 2012. "Political Terror Scale." http://www.politicalterrorscale.org/
Gibney, Mark, and Matthew Dalton. 1996. "The Political Terror Scale." In *Human Rights and Developing Countries*, ed. D. L. Cingranelli. Vol. 4 of *Policy Studies and Developing Nations.* Greenwich, CT: JAI Press, 73–84.
Gleditsch, Kristian Skrede. 2002. "Expanded Trade and GDP Data." *Journal of Conflict Resolution* 46 (5): 712–24.
Gleditsch, Kristian S., and Michael D. Ward. 1997. "Double Take: A Re-examination of Democracy and Autocracy in Modern Politics." *Journal of Conflict Resolution* 41 (3): 361–83.
Goldstein, Robert Justin. 1978. *Political Repression in Modern America, From 1870 to Present*. Cambridge, MA: G. K. Hall.
Goodman, Ryan, and Derek Jinks. 2003. "Measuring the Effects of Human Rights Treaties." *European Journal of International Law* 14 (1): 171–83.
Hafner-Burton, Emilie M. 2013. *Making Human Rights a Reality*. Princeton, NJ: Princeton University Press.
Hafner-Burton, Emilie M., and James Ron. 2009. "SEEING DOUBLE Human Rights Impact through Qualitative and Quantitative Eyes." *World Politics* 61 (2): 360–401.
Hafner-Burton, Emilie M., and Kiyoteru Tsutsui. 2005. "Human Rights in a Globalizing World: The Paradox of Empty Promises." *American Journal of Sociology* 110 (5): 1373–411.
Hafner-Burton, Emilie M., and Kiyoteru Tsutsui. 2007. "Justice Lost! The Failure of International Human Rights Law to Matter where Needed Most." *Journal of Peace Research* 44 (4): 407–25.
Harff, Barabara. 2003. "No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder since 1955." *American Political Science Review* 97 (1): 57–73.
Harff, Barbara, and Ted R. Gurr. 1988. "Toward Empirical Theory of Genocides and Politicides: Identification and Measurement of Cases Since 1945." *International Studies Quarterly* 32 (3): 359–71.
Hathaway, Oona A. 2002. "Do Human Rights Treaties Make a Difference?" *Yale Law Journal* 111 (8): 1935–2042.
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: 233–45.
Heidelberger, Philip, and Peter D. Welch. 1983. "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research* 31 (6): 1109–44.
Hill, Jr., Daniel W. 2010. "Estimating the Effects of Human Rights Treaties on State Behavior." *Journal of Politics* 72 (4): 1161–74.
Hill Jr., Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* (forthcoming).
Hill, Jr., Daniel W., Will H. Moore, and Bumba Mukherjee. 2013. ""Information Politics v Organizational Incentives: When are Amnesty International's "Naming and Shaming" Reports Biased?" *International Studies Quarterly* 57 (2): 219–32.
Hollyer, James R., and B. Peter Rosendorff. 2011. "Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance." *Quarterly Journal of Political Science* 6: 275–327.
Hopgood, Stephen. 2006. *Keepers of the Flame: Understanding Amnesty International*. Ithaca, NY: Cornell University Press.
Innes de Neufville, Judith. 1986. "Human Rights Reporting as a Policy Tool: An Examination of the State Department Country Reports." *Human Rights Quarterly* 8 (4): 681–99.
Jabine, Thomas B., and Richard P. Claude, eds. 1992. *Human Rights and Statistics: Getting the Record Straight*. Philadelphia, PA: University of Pennsylvania Press.
Jackman, Simon. 2008. "Measurement." In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier. New York: Oxford University Press.
Keck, Margaret, and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.
Keith, Linda Camp. 1999. "The United Nations International Covenant on Civil and Political Rights: Does it Make a Difference in Human Rights Behavior?" *Journal of Peace Research* 36 (1): 95–118.
Keith, Linda Camp, C. Neal Tate, and Steven C. Poe. 2009. "Is The Law A Mere Parchment Barrier To Human Rights Abuse?" *Journal of Politics* 71 (1): 644–60.
Korey, William. 2001. *NGOs and the Universal Declaration of Human Rights: A Curious Grapevine*. New York, NY: Palgrave Macmillan.
Lake, David A., and Wendy Wong. 2009. "The Politics of Networks: Interests, Power, and Human Rights Norms." In *Networked Politics*, ed. Miles Kahler. Ithaca, NY: Cornell University Press.
Landman, Todd, and Marco Larizza. 2009. "Inequality and Human Rights: Who Controls What, When, and How." *International Studies Quarterly* 53 (3): 715–36.
Lawyers Committee for Human Rights. 1997. *Critique: Review of the U.S. Department of State's Country Reports on Human Rights Practices for 1996*. Philadelphia: University of Pennsylvania Press.
Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum Associates.
Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
Lupu, Yonatan. 2013*a*. "Best Evidence: The Role of Information in Domestic Judicial Enforcement of International Human Rights Agreements." *International Organization* 67 (3): 469–503.
Lupu, Yonatan. 2013*b*. "The Informative Power of Treaty Commitment: Using the Spatial Model to Address Selection Effects." *American Journal of Political Science* 57 (4): 912–25.
Marshall, Monty G., Ted R. Gurr, and Barbara Harff. 2009. "PITF - STATE FAILURE PROBLEM SET: Internal Wars and Failures of Governance, 1955–2009." *Dataset and Coding Guidelines*.
Marshall, Monty, Keith Jaggers, and Ted R. Gurr. 2013. "Polity IV Project: Political Regime Characteristics and Transitions 1800-2010 Dataset Users' Manual." *www.systemicpeace.org/polity/polity4.htm*
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.
Mason, T. David, and Dale A. Krane. 1989. "The Political Economy of Death Squads: Toward a Theory of the Impact of State-Sanctioned Terror." *International Studies Quarterly* 33 (2): 175–98.
Moore, Will H. 1998. "Repression and Dissent: Substitution, Context, Timing." *American Journal of Political Science* 42 (3): 851–73.
Moore, Wil H. 2000. "The Repression of Dissent: A Substitution Model of Government Coercion." *Journal of Conflict Resolution* 44 (1): 107–27.
Morrow, James D. 2007. "When do States Follow the Laws of War?" *American Political Science Review* 101 (3): 559–72.
Murdie, Amanda, and Tavishi Bhasin. 2011. "Aiding and Abetting? Human Rights INGOs and Domestic Anti- Government Protest." *Journal of Conflict Resolution* 55 (2): 163–91.
Murdie, Amanda, and David R. Davis. 2012. "Looking in the Mirror: Comparing INGO Networks Across Issue Areas." *Review of International Organizations* 7 (2): 177–202.
Neumayer, Eric. 2005. "Do International Human Rights Treaties Improve Respect for Human Rights?" *Journal of Conflict Resolution* 49 (6): 925–53.
Nordås, Ragnhild, and Christian Davenport. 2013. "Fight the Youth: Youth Bulges and State Repression." *American Journal of Political Science* 57 (4): 926–40.

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.
Plummer, Martyn. 2010. "JAGS (Just Another Gibbs Sampler) 1.0.3 Universal." http://www-fis.iarc.fr/ martyn/software/jags/
Poe, Steven C. 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*, eds. S. Carey and S. Poe. Aldershott: Ashgate, 16–42.
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.
Poe, Steven C., and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis." *American Political Science Review* 88 (4): 853–72.
Poe, Steven C., C. Neal Tate, and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global Cross-National Study Covering the Years 1976–1993." *International Studies Quarterly* 43 (2): 291–313.
Poole, Keith T., and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (1): 228–78.
Poole, Keith T., and Howard Rosenthal. 1997. *A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.
Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development*. New York: Cambridge University Press.
Rasch, Georg. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.
Rummel, Rudolph J. 1994*a*. *Death by Government: Genocide and Mass Murder in the Twentieth Century*. New Brunswick, NJ: Transaction Publishers.
Rummel, Rudolph J. 1994*b*. "Power, Geocide and Mass Murder." *Journal of Peace Research* 31 (1): 1–10.
Rummel, Rudolph J. 1995. "Democracy, Power, Genocide, and Mass Murder." *Journal of Conflict Resolution* 39 (1): 3–26.
Schnakenberg, Keith E., and Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2 (1): 1–31.
Schoultz, Lars. 1981. "US Foreign Policy and Human Rights Violations in Latin America: A Comparative Analysis of Foreign Aid Distributions." *Comparative Politics* 13 (2): 149–70.
Sikkink, Kathryn. 2011. *The Justice Cascade: How Human Rights Prosecutions Are Changing World Politics*. The Norton Series in World Politics: W.W. Norton and Company.
Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *American Political Science Review* 94 (4): 819–35.
Simmons, Beth A. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge University Press.
Simmons, Beth A., and Daniel J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99 (4): 623–31.
Sinkovits, Robert S., Pietro Cicotti, Shawn Strande, Mahidhar Tatineni, Paul Rodriguez, Nicole Wolter, and Natasha Bala. 2011. "Data Intensive Analysis on the Gordon High Performance Data and Compute System." *KDD '11 Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 747–8.
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, and Steven E. Johnson. 1984. "Human Rights and US Foreign Assistance from Nixon to Carter." *Journal of Peace Research* 21 (3): 215–26.
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.
Sundberg, Ralph. 2009. "Revisiting One-sided Violence: A Global and Regional Analysis." In *States in Armed Conflict*, eds. Lotta Harbom and Ralph Sundberg. Uppsala: Universitetstryckeriet.
Taylor, Charles Lewis, and David A. Jodice. 1983. *World Handbook of Political and Social Indicators Third Edition*. Vol. 2, Political Protest and Government Change. New Haven, CT: Yale University Press.
Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52 (1): 201–17.
Trochim, William M.K., and James P. Donnelly. 2008. *Research Methods Knowledge Base*. 3rd ed. Mason, OH: Atomic Dog.
van Schuur, Wijbrandt H. 2003. "Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory." *Political Analysis* 11 (2): 139–63.
Voeten, Erik. 2000. "Clashes in the Assembly." *International Organization* 54 (2): 185–215.
Von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *American Political Science Review* 99 (4): 611–22.
Vreeland, James Raymond. 2008. "Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture." *International Organization* 62 (1): 65–101.
Wayman, Frank W., and Atsushi Tago. 2010. "Explaining the Onset of Mass Killing, 1949–87." *Journal of Peace Research* 47 (1): 3–13.
Wong, Wendy H. 2012. *Internal Affairs: How the Structure of NGOs Transforms Human Rights*. Ithaca, NY: Cornell University Press.
Wood, Reed M. 2008. "'A Hand Upon the Throat of the Nation': Economic Sanctions and State Repression, 1976–2001." *International Studies Quarterly* 52 (3): 489–513.
Zanger, Sabine C. 2000. "A Global Analysis of the Effect of Political Regime Changes on Life Integrity Violations, 1977–93." *Journal of Peace Research* 37 (2): 213–33.