Yes, Human Rights Practices Are Improving Over Time

Christopher J. Fariss*†

click here for the most up to date version of this manuscript

Abstract

To document human rights, monitoring organizations establish a standard of accountability, or a baseline set of expectations that states ought to meet in order to be considered respectful of human rights. If the standard of accountability has meaningfully changed then the categorized variables from human rights documents will mask real improvements. Cingranelli and Filippov question whether the standard of accountability is changing and if data on mass-killings are part of the same underlying conceptual process of repression as other abuses. These claims are used to justify alternative models, showing no improvement in human rights. However, by focusing on the coding process, these authors misunderstand that the standard of accountability is about how monitoring organizations produce documents in the first place and not how academics use published documents to create data. Simulations and latent variables that model time in a substantively meaningful way validate the conclusion that human rights are improving.

*Assistant Professor, Department of Political Science, University of Michigan, cjfariss@umich.edu; cjf0006@gmail.com

†Special thanks goes to Michael Kenwick and Kevin Reuning, who have provide an immeasurable amount of assistance and support as I prepared this response. Much of the work builds on several joint measurement projects that are currently underway (Reuning, Kenwick and Fariss, 2018). I would also like to acknowledge research support from the SSK (SocialScience Korea) Human Rights Forum, the Ministry of Education of the Republic of Korea, and the National Research Foundation of Korea (NRF- 2016S1A3A2925085).

1 Introduction

The standard of accountability is the set of expectations developed by human rights monitoring organizations about the specific responsibilities that governments around the world have, and *ought* to meet, with respect to the treatment of individuals. It is also the core concept from a theory about how the organizational structures and procedures of human rights monitoring organizations produce information about state behaviors over time (Fariss, 2014). In short, the standard of accountability continues to evolve as activists, lawyers, jurists, norm entrepreneurs, regional human rights courts, NGOs, IGOs, government agents, and other actors call attention to state behaviors, create innovative legal arguments, and build new institutions designed to protect the rights of individuals (e.g., Brysk, 1994; Clark, 2001; Dancy, 2016; Dancy and Fariss, 2017; Dancy and Michel, 2015; Mayerfeld, 2016; Sikkink, 2011, 2017). If the standard of accountability has meaningfully changed over time then the categorized variables from human rights documents will mask real improvements in human rights over time (Fariss, 2014, 2018*a*,*b*).

In a recent critique, Cingranelli and Filippov (2018) question whether the standard of accountability is changing and if data on mass-killings are part of the same underlying conceptual process of repression as other human rights abuses like torture and political imprisonment. These authors base their argument on a mischaracterization of the theory from Fariss (2014). Summarizing the theory, Cingranelli and Filippov (2018) state that "[h]uman rights scores may be inconsistent over time, because: (a) human rights reports have gotten longer, and more information, by itself, may have influenced coders to assign lower scores; (b) coders may have applied more stringent standards in more recent years; and (c) there may be new types of critiques included in more recent reports" (pg. 3). By focusing on how political scientists code documents, Cingranelli and Filippov (2018) misunderstand that the standard of accountability is about the original documentation process by monitoring organizations and not the academic coding process.[1]

The theoretical distinction between actors (document producers vs. academic coders) is important for making conceptual distinctions between different indicators of repression and when introducing modifications to latent variable models. Building on the mischaracterization of the theory of standard of accountability, Cingranelli and Filippov (2018) make two related claims that they then use to justify two

[1] As Clark and Sikkink (2013) argue, academic coders may be influenced by larger quantities and greater quality of information when coding reports. Though the argument in Fariss (2014) builds on this idea of an information paradox, it shifts the conceptual focus from the academic coders to the report-producers.

alternative latent variable modeling decisions. First, the coding process for all existing human rights data are potentially affected by the standard of accountability. Second, data on mass-killings are not part of the same underlying concept of repression as other human rights abuses like torture and political imprisonment and should be considered as separate conceptual dimensions. Cingranelli and Filippov (2018) use these claims to suggest that the latent variable models of human rights presented by Fariss (2014) are misspecified. Then, based on one alternative latent variable model specification that includes all variables and one that includes only standards-based variables, Cingranelli and Filippov (2018) use estimates from these two models to conclude that human rights are not improving over time.

In response to these claims, I first discuss the critique of the standard of accountability as it relates to the documentation process conducted by monitoring organizations that generates qualitative reports and the coding process conducted by academics that generates categorical data from those reports. I also discuss the use of evidence from data on mass-killings in conjunction with data on other forms of human rights abuse. These discussions are important because they form the theoretical justification for the different latent variable model specifications presented in Fariss (2014) and here. Second, I present a simulation exercise, which illustrates the identification problem in a latent variable model where all of the item difficulty parameters vary from year to year (instead of estimating a single item difficulty parameter per item for all units). The evidence from the simulations demonstrates that the model suggested by Cingranelli and Filippov (2018) is not identified with respect to time because it resets the yearly average of the latent variable to 0 no matter the values of the data. Relatedly, randomly generated data, as presented by Cingranelli and Filippov (2018), does not meaningfully change latent variable model estimates because the simulated values only add random noise to the models, which average out to 0 each year. Third, I present an updated version of the latent variable model of human rights and use construct validity and posterior predictive evidence to compare the two original models presented in Fariss (2014) and the alternative version presented by Cingranelli and Filippov (2018). These validity assessments show the substantive consequences of not accounting for time in a substantively meaningful way. Fourth, I present yearly estimates from several additional latent variables models, which are based on different subsets of the available human rights variables. Nearly all models, with the exception of models that contain information about torture derived from the US State Department reports, show an

improving trend in human rights over time (these models are each estimated with single item difficulty parameters for each item, which is the same as the constant standard model presented in Fariss (2014)). These model comparisons also reveal which variables are driving the differences between estimates from the changing standard of accountability model and its alternatives and provides new inferences that I hope will spur additional theorizing and measurement modeling. These new results are also consistent with new expert-coded human rights data from VDEM (Coppedge et al., 2014; Pemstein, Tzelgov and Wang, 2015), which provides convergent validity evidence in support of the conclusion that human rights are improving over time. Overall, what this evidence demonstrates is not that the improving yearly average of human rights in (Fariss, 2014) is being driven by large-scale event-based indicators, but rather, that certain standards-based variables are masking the yearly improvements across many indicators (see the Appendix) and latent variables estimates because monitors are increasingly likely to observe and report insistences of torture and ill treatment in more recent years. Finally, a supplementary appendix and line by line response address every point from Cingranelli and Filippov (2018) in detail.

2 Conceptual Foundations for Human Rights Latent Variables

2.1 The Documentation Process and Coding Process of Human Rights

The goal of measurement is to define an operational procedure that takes information and creates data free from conceptual (translational) error and measurement error (Adcock and Collier, 2001; Fariss and Dancy, 2017). A categorization process like the Political Terror Scale (PTS) (Gibney et al., 2017) or the CIRI human rights project (Cingranelli, Richards and Clay, 2015) are operational procedures designed to be consistently applied to human rights documents in order to categorize aggregated country-year human rights practices. Fortunately for these academic teams, large-scale monitoring efforts systematically produce yearly human rights reports that are publicly available and cover nearly every country in the world. The political science teams that work to categorize the information contained in these human rights reports can take advantage of the fact that these reports are produced using a standardized process each year. This is what the human rights community means when they refer to the PTS and CIRI variables as standards-based.Conceptually, it is important to note that these standards are not in reference to the

coding procedures used to code the data but rather the standardized procedures used by the monitoring organizations to produce the human rights reports each year. It is this standardized information which is then used by the political science teams to categorize information into human rights data. But what if the standards used to produce the primary source human rights documents change?

The theory of the standard of accountability helps answer this question because it is focused on the organizational structures and procedures that are developed and implemented to document human rights abuses by monitoring organizations like Amnesty International and the US State Department. The PTS and CIRI teams have very little input into these organizational processes. The categorical indicators coded from the reports are manifest of a complex process that begins with the human rights abuses themselves, the observation, collection, and corroboration of allegations about those human rights abuses, the organization of those allegations into a structured narrative account contained within the country reports, and finally the coding/categorization process of that content by academics. With sufficient understanding of this process, the latent variable model provides estimates of the relative level of human rights respect for each country-year unit in the sample based on the categorical values of the reports and theoretical knowledge and information about the rest of the underlying process. To compare categorical values, the CIRI and PTS teams must assume that the processes that build up to the publication of the human rights reports are constant for each report, in every year and for every country, and that variation in the content of each report is only attributable to differences in the underlying human rights condition for each country-year unit. Though the categorized values coded from the human rights reports by PTS and CIRI reliably reflect the content of the human rights reports, these values are not able to directly capture any differences in the processes that lead up to their publication.

The theoretical distinction between actors (document producers vs. academic coders) is important for making conceptual distinctions between different indicators of repression and when introducing modifications to latent variable models. In order to justify the specification of their alternative latent variable model, Cingranelli and Filippov (2018) suggest that the changing standard of accountability can affect the coding process of the events-based variables. Specifically, they suggest that "[t]here is a higher likelihood now that mass killings in remote places will be recorded. Coding rules for recording mass killings may be changing. Coders may have applied more stringent standards in more recent years. And coding rules

across mass killing recording projects may be be becoming more or less consistent with one another" (pg.5).

To review, these variables cover specific forms of repression, mostly related to state sanctioned killing: mass killings, mass repression, genocide, politicide, executions, negative-sanctions, or one sided government killings. The standard of accountability likely affects the documentation used to code these variables as well. However, unlike the CIRI, PTS, Hathaway, and ITT data projects, the event-based variables are not direct categorizations of documents but rather, are binary indicators that are coded 1 if sufficient documentary information exists in the historical record to support such a categorization. For the standards-based variables, the documents are directly categorized. Because the documents are never updated or revised, the standards-based variables are rarely updated. For the event-based variables, documentary evidence is taken from multiple sources and used to look for evidence that a particular type of repressive event occurred. If new documentary evidence emerges about a specific type of repressive event, the categorized value for the country-year unit is updated. Thus, these are fundamentally distinct categorization processes. The first categorization process relies exclusively on the content from the individual country-year report. The second coding process instead relies on a set of documents and, for many of the variables considered in this paper, is updated and repeated when new information enters the historical record. The event-based categorization process is therefore able to potentially address variation in the underlying documentation processes that generates information because these variables are each based on set of different documents and are updated periodically. The standards-based coding process cannot directly account for this variation.

To clarify, Fariss (2014) does not argue that mass killing are recorded more accurately than other forms of violations in any particular documentary source, when that documentary source is produced. Rather, the distinction between standards-based and events-based variables in Fariss (2014) is about how the documentary evidence is used to create categorical data. Specifically, Fariss (2014) distinguishes between the direct categorization of the documentary evidence (standards-based variables) and the broader use of documents to find evidence of a specific type of event. It need not be the case that large scale events are recorded more accurately in the historical record than other violations because evidence for many repressive events does not necessarily enter the historical record as they are occurring (Davenport

and Ball, 2002; Krüger et al., 2013). It is therefore important to continue to update the historical source material used to create event-based categorical variables as all of the events-based coding teams have done a various points in time (e.g., Eck and Hultman, 2007; Harff, 2003; Harff and Gurr, 1988; Rummel, 1994; Taylor and Hudson, 1972; Taylor and Jodice, 1983). The difference between the specification of the constant standard model and the changing standard model presented in Fariss (2014) is based on the difference in how these two types of variables are categorized. The standard of accountability is likely changing all of the documentary evidence used by the different coding teams but the event-based variables are categorized using many sources and updated over time, which helps to account for bias from particular sources. This makes the event-based variables suitable to act as a baseline for comparison with the standards-based variables that do not share this feature.

2.2 The Concept of Physical Integrity Rights: One or Two Dimensions?

Cingranelli and Filippov (2018) also suggest that large scale killing events are a distinct repertoire of state-sanctioned repression in comparison to other forms of physical integrity abuses such as political imprisonment, and ill treatment and torture because governments adopt different tactics to implement these polices. As such, they should be treated as independent dimensions in analysis. However, these types of repressive practices are conceptually and empirically related to state-sanctioned practices that are associated with disappearances, extra-judicial killings, and the large scale occurrence of killings as well. This theoretical understanding forms the basis for the primary conceptual definition of "repression" or violations of "physical integrity rights" in the literature, which 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 (Davenport, 2007; Goldstein, 1978).

This argument by Cingranelli and Filippov (2018) is similar to an argument made by McCormick and Mitchell (1997). McCormick and Mitchell (1997) argued that physical integrity rights should be considered along two dimensions: killing and disappearances, which often end in death, and torture and political imprisonment, which are about the treatment of the living. In the article that introduced the four CIRI physical integrity variables, Cingranelli and Richards (1999) argued against the 2-dimensional

conceptualization and used a statistic from Mokken (1971) to demonstrate the indicators of these four types of physical integrity violations scale together along one dimension.

The scaling result from Cingranelli and Richards (1999) is supported by additional evidence about the relationship between all of the CIRI variables presented by Fariss and Schnakenberg (2014), which shows a high degree of complementarity and no evidence of substitution between the four physical integrity rights at the aggregate country-year level, and in Schnakenberg and Fariss (2014), which provided additional validation of the scaleability of the four physical integrity variables. Finally, Fariss (2014) discusses a multidimensional IRT model in the supplementary appendix of that article and finds no empirical support for a second dimension from the 13 indicators considered in that article. In summary, contrary to the argument from McCormick and Mitchell (1997) and now from Cingranelli and Filippov (2018), there is no empirical evidence from Cingranelli and Richards (1999) or more recently from Fariss and Schnakenberg (2014), or Schnakenberg and Fariss (2014), that repressive tactics scale on more than one dimension at the country-year level of aggregation. This does not mean that, below the country-year level of aggregation, substitution between repressive practices might be occurring or that different tactical choices are associated with specific forms of repression or violence, which is consistent with the argument and evidence from violence in Colombia from Guitiérrez-Sanín and Wood (2017). It only means that empirically, these physical integrity variables are conceptually related and are useful for scaling and comparing country-year units.

It is important to further consider the aggregation of government policies or tactics as they relate to the use and observability of different forms of repression. Substantively, there is substantial evidence that certain forms of human rights abuses, what Cingranelli and Filippov (2018) label as "lesser forms of abuse" were under-reported in earlier periods, particularly during periods when more egregious forms of abuse were prevalent. If the presence of one repressive tactic reduces the probability that another tactic is observed by a monitoring organization or dampens the retribution faced by a leader caught using the tactic, those tactics may be complements, which makes observing each type of abuse difficult as the scale of other abuses increases (Fariss and Schnakenberg, 2014). This first idea is consistent with Brysk (1994), who argues that "[i]ncidents of kidnapping and torture which would register as human violations elsewhere did not count in Argentina. The volume of worse rights abuses set a perverse benchmark and

absorbed monitoring capabilities" (pg 681).

This logic implicates how the documentation of abuses occurs in highly respectful cases as well. For example, the CIRI human rights project codes the highly transparent case of Sweden as a state that uses ill treatment and torture every year since 2005 (Eck and Fariss, 2018). What the insights from Brysk (1994) and Eck and Fariss (2018) reveal, is that certain forms of abuse are relatively easier to observe when the overall level of human rights abuse is low but relatively more difficult to observe when the overall level of human rights abuse is high. This is because monitoring capacity is not limitless but is increasingly effective as the volume of abuses decreases. This seems to be especially the case for instances of ill treatment and torture.

As a tactic, torture and ill treatment may be intended to extract information from some individuals, whereas it may be used to the intimidation of others. Though the goals of the use of torture and ill treatment, or the ability of the state to structure institutions that completely eliminate the practice, likely varies within different states, the overall aggregation of information about the practice at the country-year level is empirically related to other instances of physical integrity abuse (Brysk, 1994; Eck and Fariss, 2018). Thus, observationally, it is more difficult for monitoring organizations to detect torture and ill treatment in comparison to the other forms of physical integrity rights abuse as the scale of abuses increases. In the new latent variable models below, I demonstrate that, consistent with this logic, information on torture and ill-treatment derived from the State Department reports is the most sensitive to the changing standard of accountability relative to the other event-based data but also all of the other standards-based categorical indicators as well.

3 Simulation Analysis: Constant Difficulty Parameters Compared to Temporally Varying Difficulty Parameters

Cingranelli and Filippov (2018) charge that Fariss (2014) failed to assess the assumptions of the two latent variables models. To support this claim, Cingranelli and Filippov (2018) create random ordered categorical variables and re-estimate the latent variable model with these additional variables in place of the standards-based variables. They wish to draw the inference that the standards-based variables do

not provide meaningful information in the model by comparing the yearly mean point estimates between these two models. But this inference is not valid. Because the simulated data are not generated from any underlying model related to the other data, Cingranelli and Filippov (2018) have essentially just added random noise to the estimates. Unfortunately, there is nothing we can learn from this simulation other than what the unchanged events-based variables already show us. This is because, though adding random data to the units in the latent variable model will randomly shift the position of some units, it will not change the average for these units in each year.

Relatedly, in their abstract, Cingranelli and Filippov (2018) state that I rely on "stringent assumptions" to argue that I "heavily weighted rare incidents of mass killings such as genocide". However, this is untrue. The latent variable models in Fariss (2014) incorporate the event-based variables in exactly the same way across the different specifications. The standards-based variables are treated differently, but it is not through the item-weights, but rather through their item difficulty parameters. The term item-weights is usually used to describe the item discrimination parameters in the IRT models, which are analogous to slope parameters in a logit or ordered-logit function, whereas the item difficulty parameters are analogous to intercepts or cut-points. I treat standards-based variables differently in accordance with the theory of a changing standard of accountability, which I discuss in great detail above and in Fariss (2014). Readers should be aware that the assertion made by Cingranelli and Filippov (2018) is misleading for two reasons: first, because they do not clearly discuss which of the parameters they are criticizing, and second, because they provide no empirical evidence to support their claim about the treatment of the item-weights in Fariss (2014).

The country time-series plots that Cingranelli and Filippov (2018) provide are discussed without any systematic statistical analysis. Contrary to their general claim, the evidence in these graphs supports the use of both the events-based and standards-based data to measure the same theoretical concept of physical integrity because they are highly related even though they are estimates from different sets of human rights variables. Thus, each of the standards-based variables provide meaningful information for the placement of each of the country-year units relative to all the others (see the Appendix for more details about this point). This is evidence of convergent validity, a specific type of construct validity, which I discuss in the results section below. The rest of the evidence for the Cingranelli and Filippov

(2018) critique is based on estimates from two alternative latent variable models that show that human rights have not improved over time. However the first of these alternative models is not identified with respect to time, which makes the latent variable estimates from this model not comparable from year to year (though country comparisons within each year are possible). I demonstrate this issue with a simple simulation.

For the simulation, I set the number of units $N = 30$, the number of time periods $T = 10$, and the number of binary items $K = 5$. The simulation takes an initial draw from θ, which is the latent variable for each unit: $\theta_{t,1} \sim \text{N}(-2, 1)$. It then takes the remaining draws from $\theta_{i,t} \sim \theta_{n,t-1} + c_t + \text{N}(0, \sigma)$. Where c_t is a constant value for each time period set so that the average value of the θ_t increases by $\frac{4}{9}$ over the 10 time periods, which beings at -2 and ends at 2. The innovation standard deviation is set to $\sigma = 0.05$, which approximates this parameter from the latent human right variable (Schnakenberg and Fariss, 2014). The $K = 5$ item difficulty parameters are set to $\alpha_k = 0$ and the item discrimination parameters are set to $\beta_k = 1$. The following data generating process generates the simulated items for each unit: $y_{i,t,k} \sim \text{Bernoulli}\left(\Phi(2(\alpha_k + \beta_k \theta_{i,t}))\right)$ where Φ is the cumulative distribution function of the standard normal distribution.[2]

Next, I estimate posterior distributions for the latent variables with the simulated data in six different models. The first model is a dynamic latent variable model with a single, fixed or constant difficulty parameter that is estimated for each of the items (α_k). This model matches the constant standard model from Fariss (2014). The other models are also dynamic but with a set of difficulty parameters ($\alpha_{t,k}$) which are allowed to vary for each of the 10 time periods. Specifically, for these models, $\alpha_{t,k}$ parameter is estimated for each of the 10 time periods for 1, 2, 3, 4, or all 5 items, while only a single α_k parameter is estimated for the remaining item(s) (i.e., one constant α_k parameter is estimated for 4, 3, 2, 1, or 0 items respectively). The model that estimates the varying difficulty parameters for all 5 items (0 constant), matches the setup of the model proposed by Cingranelli and Filippov (2018) (all-varying standard model). The models in between are similar to the changing standard of accountability model (Fariss, 2014).

In Figure 1, the dark points represent the true mean for each time period from the simulated data. The grey bars represent the posterior distribution for each time period mean estimated from each of the different latent variable models. The model in the top left of Figure 1 is estimated with constant

[2] See similar simulations in Reuning, Kenwick and Fariss (2018).

item difficulty parameters, which is analogous to a standard dynamic latent variable model (dynamic with respect to the latent trait for each unit but constant or fixed with respect to each item difficulty parameter). The model on the lower right side is estimated with time specific item difficulty parameters for all items, which is analogous to the model proposed by Cingranelli and Filippov (2018). This model is not identified with respect to time because it centers the mean for the units in each time period at 0, which makes it impossible to make over time comparisons of the latent variables estimated from the model on the right. Any latent variable model that allows all the item difficulty parameters to vary over time will behave in this way. As long as the item-difficulty parameter for at least one item is constant (i.e., only one item difficulty parameter is estimated for all units), the model recovers evidence for the change over time. Note that in this simulation, the observed data are increasing on average because they are generated from a latent trait that is increasing over the 10 time periods, which is not the case for the standards-based human rights data presented in Fariss (2014).

Figure 2 demonstrates the ability of the models to estimate the latent mean across time periods as the number of constant item difficulty parameters decreases to 0 (the all-varying model proposed by Cingranelli and Filippov (2018)). Like the other four models displayed in Figure 1, the latent variable model that incorporates the changing standard of accountability is specified by allowing some of the item difficulty parameters to change over time but not all of them. Unlike the all-varying difficulty parameter model that Cingranelli and Filippov (2018) propose however, such a model is identifiable with respect to time because it only allows some of the item difficulty parameters to vary over time while keeping approximately half of these parameters fixed or constant. In other words, the changing standard of accountability can only be accounted for when information that it influences is assessed in relation to information that is generated consistently over time. The model does this by assessing the yearly frequency of some of the standards-based variables to the overall frequency of the events-based variables and some of the other standards-based variables.

In summary, the simulation analysis demonstrates why the model proposed by Cingranelli and Filippov (2018) is not identified: what is happening with the all-varying intercept (cut-point) model, is that the mean of the latent trait is being force to 0 every year no matter what the values of the items are. So as conditions improve elsewhere, the model, by necessity, is forced to push the bad cases further down into

the latent space so that they are far enough away from the mediocre and good cases, conditional on the available data. The prior position of each unit is informative year to year but the model still must center the distribution for of all units in a given year over 0. This is because Cingranelli and Filippov (2018) have estimated a model that estimates the intercepts (cut-points) for each item for each year. There is essentially no baseline year to year so the model has to revert to the mean 0 assumption of the prior distribution. The model estimates the best intercepts (cut-points) for the data available for that item in a given year and then tries to find the best values of the latent trait in that year. Instead of letting the model use this distribution to arrange all of the units for all years, the all-varying cut-point model is resetting the distribution each year. For Cingranelli and Filippov (2018) to argue that human rights practices are not changing over time, then they need to argue for the constant standard model presented in Fariss (2014) so that the estimates are comparable from year to year. As it stands now, their model shows no improvement in human rights by assumption.

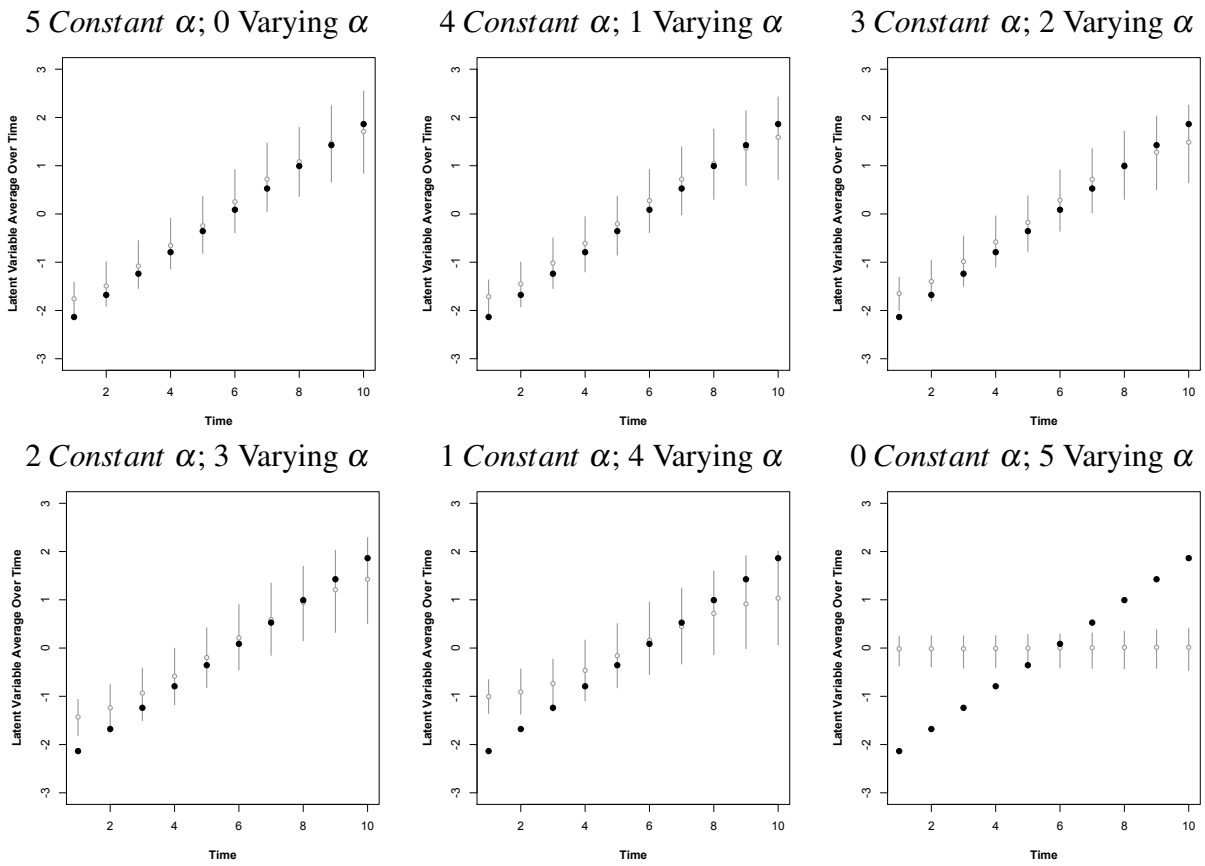

Figure 1: The dark points represent the true mean for each time period from the simulated data. The grey bars represent the posterior distribution for each time period mean from the models. Estimates are all obtained from a dynamic latent variable model with item difficulty parameters either estimated for all years together (fixed or constant) or estimated for each year individually (varying). The model on the lower right side is estimated with a time specific item difficulty parameter for every item. This model is not able to estimate the over time change because there are no observed indicators used to relate the latent estimates from year to year. The model is therefore forced to center the mean for the units in each time period at 0. All of the other models are identified with respect to time, which makes it possible to make over time comparisons of the latent variables estimated from these other models. Though ideally, increasing the number of fixed difficulty-parameters relative to varying difficulty parameters is useful because it increases the amount of information used to relate the estimates of the latent trait across time periods.

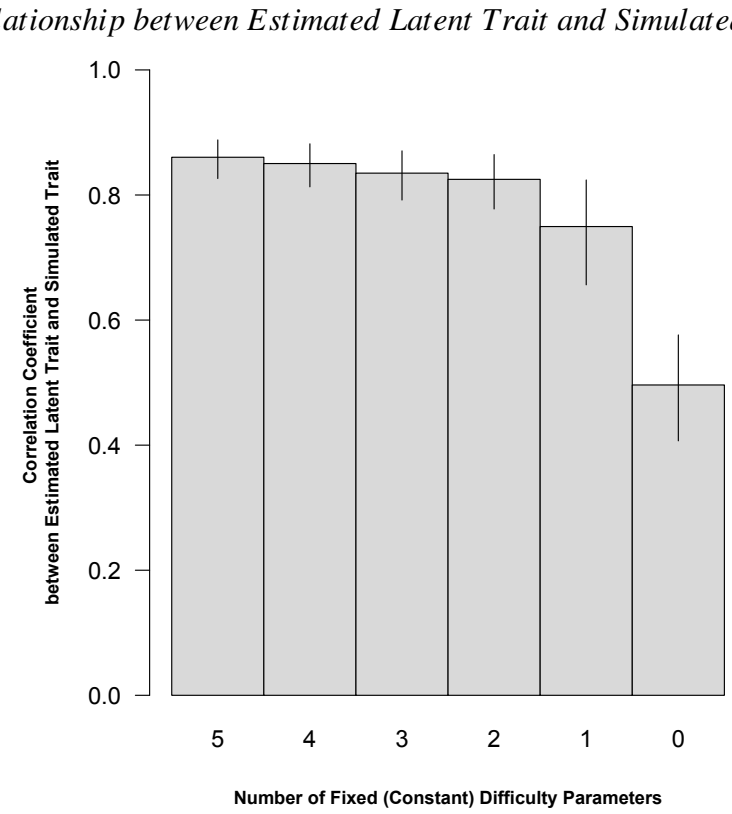

Figure 2: Correlation coefficients between estimates that are obtained from a dynamic latent variable model with item difficulty parameters that either estimated for all years together (fixed or constant) or estimated for each year individually (varying). Each bar corresponds to each of the models presented in Figure 1. The model on the right most side is estimated with a time specific item difficulty parameter for every observed item and is similar to the model suggested by Cingranelli and Filippov (2018). This model is not able to estimate the over time change because there are no observed indicators used to relate the latent estimates from year to year. This is why the correlation coefficient for this model is the smallest relative to all of the other models to which it is compared. Increasing the number of constant difficulty-parameters relative to the number of varying difficulty parameters is useful because it increases the amount of information used to relate the estimates of the latent trait across time periods. This is why the correlation between the simulated latent trait and the estimated latent trait increases as the number of constant parameters increases for the simulated data from right to left.

4 Comparing Estimates from Three Latent Variable Models

To estimate the latent variable model, each item or categorized human rights variable is linked to the latent trait — the relative level of human rights respect of one country-unit relative to all the others — using a generalized linear function (logit or ordered-logit depending on the particular variable). Mechanically, the latent variable model simply places each of the country-year units relative to one another along a single interval-level dimension with a score of 0 acting as the global average for all country-year units. All country-year units are placed relative to this average. The model proposed by Cingranelli and Filippov (2018) also places each of the country-year units relative to one another along a single interval-level dimension with a score of 0 acting as the mean for each year, which contrasts with the other latent variable models that have a global mean for units across all years. This choice in model specification makes it so that the units are not comparable between years. The all varying standard model produces a flat trend line which, by coincidence is similar, but not perfectly so, to the trend to the constant standard model from Fariss (2014). This similarity occurs because, in the all-varying standard model, it is not possible to estimate a change over time and, in the constant standard model, there is not a change over time due to the influence of the standards-based variables. Table 1 summarizes the specification of parameters for each of the three competing models. Figure 3 displays the distribution of the latent variable point estimates for the three models.

Table 1: Summary of prior distributions for latent variable and model level parameter estimates.

Parameters	**all-varying Standard**	**Constant Standard**	**Changing Standard**
country-year latent variable (first year)	$\theta_{i1} \sim N(0,1)$	$\theta_{i1} \sim N(0,1)$	$\theta_{i1} \sim N(0,1)$
country-year latent variable (other years)	$\theta_{it} \sim N(\theta_{it-1}, \sigma)$	$\theta_{it} \sim N(\theta_{it-1}, \sigma)$	$\theta_{it} \sim N(\theta_{it-1}, \sigma)$
uncertainty of latent variable	$\sigma \sim U(0,1)$	$\sigma \sim U(0,1)$	$\sigma \sim U(0,1)$
event-based item cut-points (constant)	— — — — —	$\alpha_{jk} \sim N(0,4)$	$\alpha_{jk} \sim N(0,4)$
event-based item cut-points (first year)	$\alpha_{1jk} \sim N(0,4)$	— — — — —	— — — — —
event-based item cut-points (other years)	$\alpha_{tjk} \sim N(\alpha_{t-1,jk}, 4)$	— — — — —	— — — — —
standards-based item cut-points (constant)	— — — — —	$\alpha_{jk} \sim N(0,4)$	— — — — —
standards-based item cut-points (first year)	$\alpha_{1jk} \sim N(0,4)$	— — — — —	$\alpha_{1jk} \sim N(0,4)$
standards-based item cut-points (other years)	$\alpha_{tjk} \sim N(\alpha_{t-1,jk}, 4)$	— — — — —	$\alpha_{tjk} \sim N(\alpha_{t-1,jk}, 4)$
slope	$\beta_j \sim Gamma(4,3)$	$\beta_j \sim Gamma(4,3)$	$\beta_j \sim Gamma(4,3)$
	Cingranelli and Fillippov	Fariss (2014)	Fariss (2014)

See Fariss (2014) for additional details about the specification of the human rights latent variable model.

Distribution of Mean Estimate from the Latent Variable Models

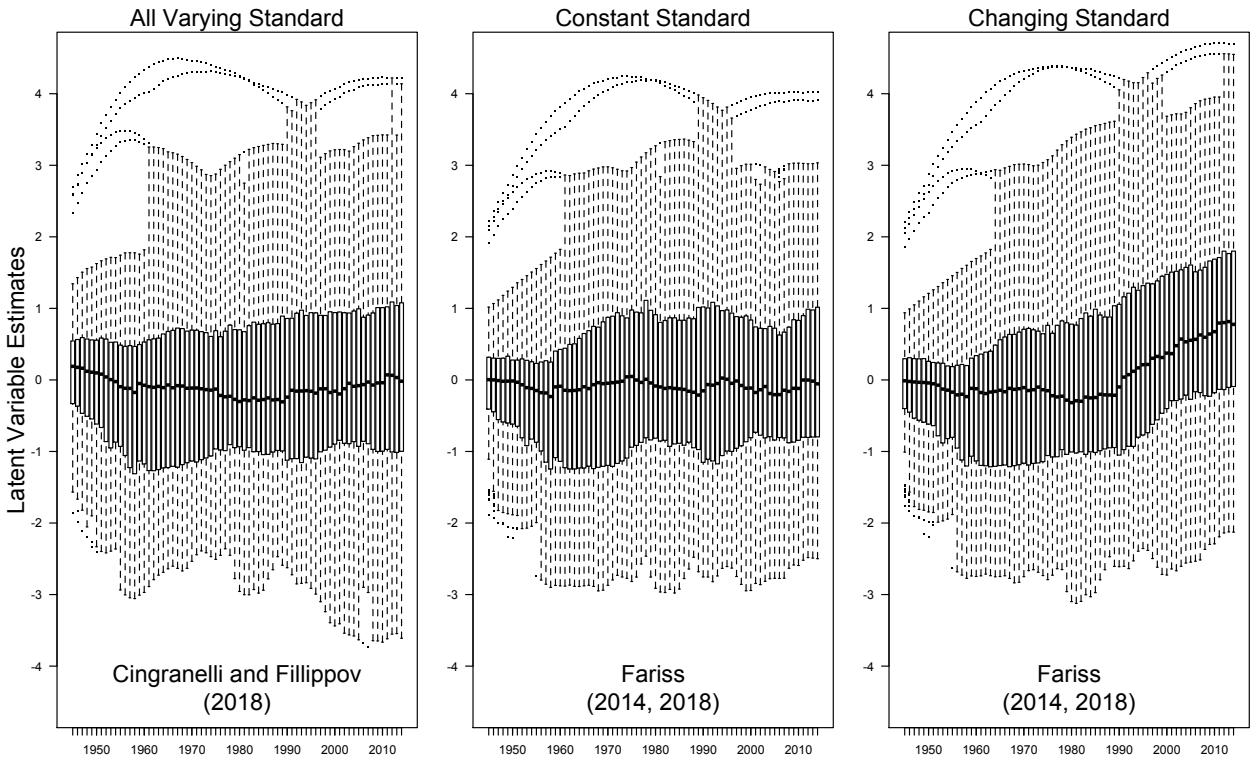

Figure 3: Each panel shows the distribution of point estimates for the human rights latent variable estimates from three competing models. The interquartile range is contained within the boxes with the median value at the center line. The dashed lines represented the values beyond the interquartile range (these plots do not incorporate uncertainty). The model proposed by Cingranelli and Filippov (2018) (left) produces a very similar trend to the constant standard model from Fariss (2014) (middle) because in the Cingranelli and Filippov (2018) model it is not possible to estimate a change over time. In the constant standard model there is not a change over time because of the influence of the standards-based variables. Because the model proposed by Cingranelli and Filippov (2018) resets the mean estimate for all the countries each year, the mean estimate can never move away from 0. This is why the median and interquartile range estimates are so consistent from year to year in the left panel. Small changes from year to year are due to new states that enter the dataset in later years.

4.1 Correlation between Latent Variable Estimates

Cingranelli and Filippov (2018) argue that the event-based variables are responsible for most of the variation in the latent variable estimates. However, the standards-based variables provide more information than the event-based variables. Moreover, all of the latent variable models, even the ones based on only some of the observed variables are highly related to one another. To see this, consider the correlational evidence presented in Figure 4, which demonstrate a very high level of agreement between estimates. This is because most of the variation in the human rights indicators is cross-sectional. However, key differences emerge when considering temporal variation as the simulation from above reveal. Cingranelli and Filippov (2018) present only one of these relationships in their paper, reporting the squared correlation from a bivariate linear regression (pg. 4). They use this statistic to suggest that the estimates from the latent variable generated from only the event-based variables explains most of the variance in the latent variable from the model that uses all the variables. However, this is not what the correlation coefficient here reveals because each of the different latent variables are not independent of one another (even when they are estimated with non-overlapping sets of human rights variables). The high correlation between the different latent variable estimates occur because each of the indicators is manifest of the same underlying concept. Though not reported by Cingranelli and Filippov (2018), the correlation coefficient is smaller than the same statistic from a latent variable model that is based on just the standards-based variables and the latent variable based on all of the items. This is because there are more standards-based variables than event-based variables and because there are more categories for each of the standards-based variables than for the events-based variables, which all happen to be binary. Because the event-based variables and standards-based variables are both capturing evidence of the same underlying physical integrity concept, they are all related and highly correlated. Even constant standard models that use only events-based variable or standards-based variables (but not both) are highly correlated, which again provides evidence that all of these variables are tapping into the same underlying theoretical concept. Note that the latent variables estimates from 1946-1975 are not from an extrapolation or interpolation, which I discuss in more detail in the Appendix.

Correlation between Latent Variables Estimates from 5 Models

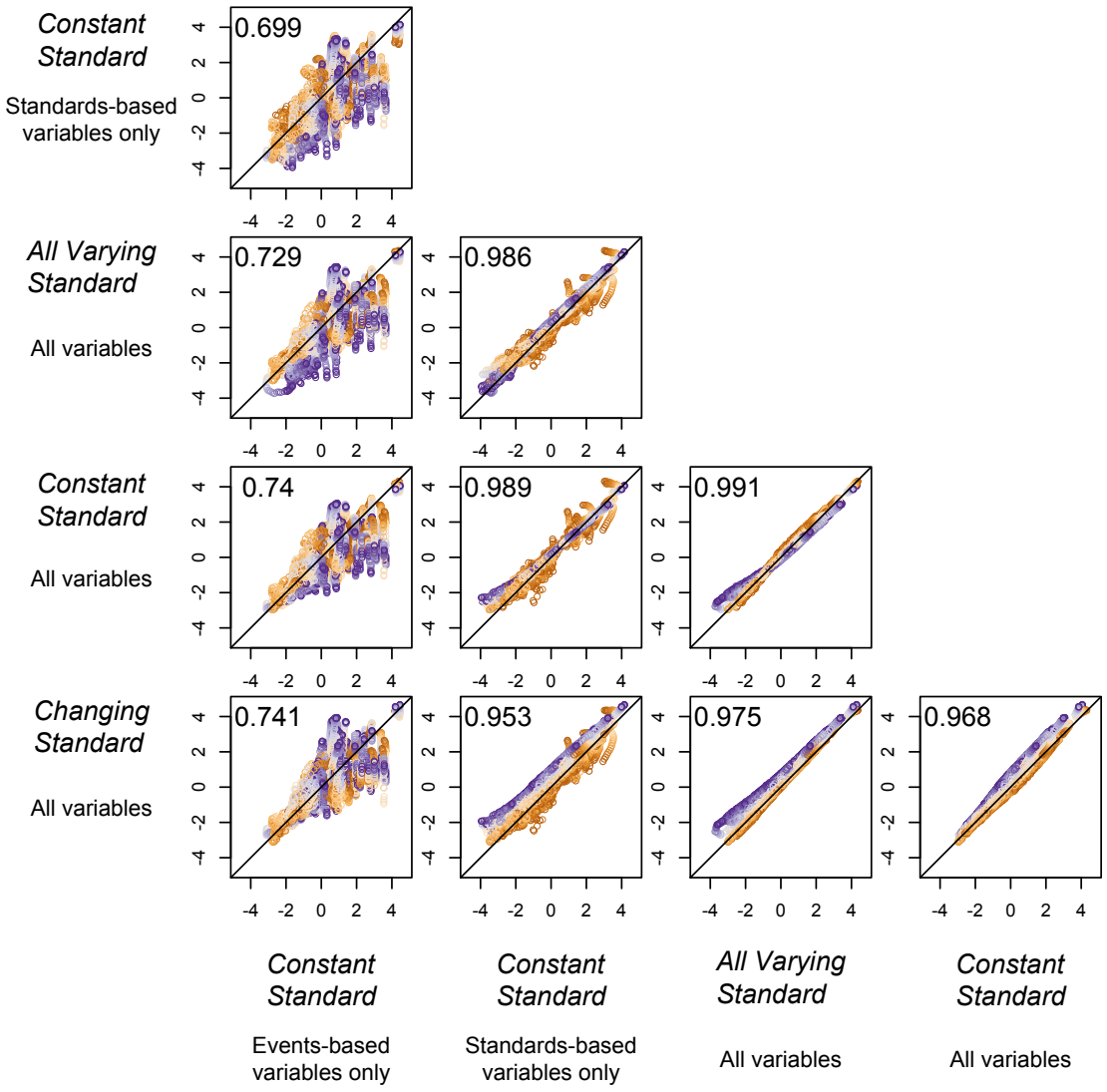

Figure 4: Correlations between five different latent variable estimates reveals a high level of agreement between the different model specifications including models with only events-based or standards-based variables. The x-axis and y-xaxis are the latent variable estimates from each the row and column latent variables estimates. Though substantively meaningful differences exists between time periods for these estimates, the high level of agreement between different estimates indicates that each latent variable estimate is tapping into the same underlying concept of physical integrity abuse. Even constant standard models that use only events-based variable or standards-based variables (but not) both are highly correlated, which again provides evidence that all of these variables are tapping into the same underlying concept. Note that, for the period 1946-1975, the correlation between the latent variable point estimates based on just the event-based item and the latent variable point estimates based on all of the items is approximately 1 because there are not standards-based variables as part of the model until 1976.

4.2 Identifying The Top-1% Worst Cases

To assess the validity of the three competing models (all varying standard, constant standard, and changing standard), I consider the ability of the latent variable estimates from each model to categorize the worst 1% of country-year cases each year. That is, in which years, do the worst cases of abuse occur? Figure 5 show that the model proposed by Cingranelli and Filippov (2018) suggests that the worst cases of human rights abuse are happening in the most recent years for which we have data, while both the constant standard model and changing standard of accountability model suggest that earlier decades contain the worst cases of abuse.

The top 20 worst cases for the Cingranelli and Filippov (2018) model are Sudan 1999-2015; and Syria 2013-2015. These are indeed cases with poor human rights practices, however, contrast these cases with those identified by the changing standard of accountability model: China 1968-1971; Uganda 1976; Afghanistan; 1980-1987; Sudan 1959, 1965-1966; and Iran 1982-1985. The constant standard model and changing standard model are in close agreement about which cases are in the top 1% worst because they are specified with respect to time. A binary indicator for whether or not the case falls in this worst case category for these two models correlates at about 0.85. The same indicator for the all-varying model correlates with either the constant standard model or changing standard model at about 0.50 or 0.43 respectively. The all-varying standard model leads to an inference that the worst levels of human rights abuse have just occurred. Though a careful analysis of each case is beyond the scope of this article, it is important to highlight that the cases selected by the Cingranelli and Filippov (2018) model are artifacts of the resetting of the mean to 0 for each year of estimates. The worst cases today are forced further down into the negative portion of the latent variable space because these positions are not estimated relative to the units in previous years but only the units in the same year. If there are more good or mediocre cases in a given year, then the bad cases need to be placed further away from the mean 0 to give space to these other cases. To suggest that the latent variable model estimates from Cingranelli and Filippov (2018) are more valid, one need also accept that the worst instances of human rights abuse are occurring today rather than the 1960s, 1970s, or 1980s.

Concurrent Validity Assessment of the 1% Worst Cases of Human Rights Abuse

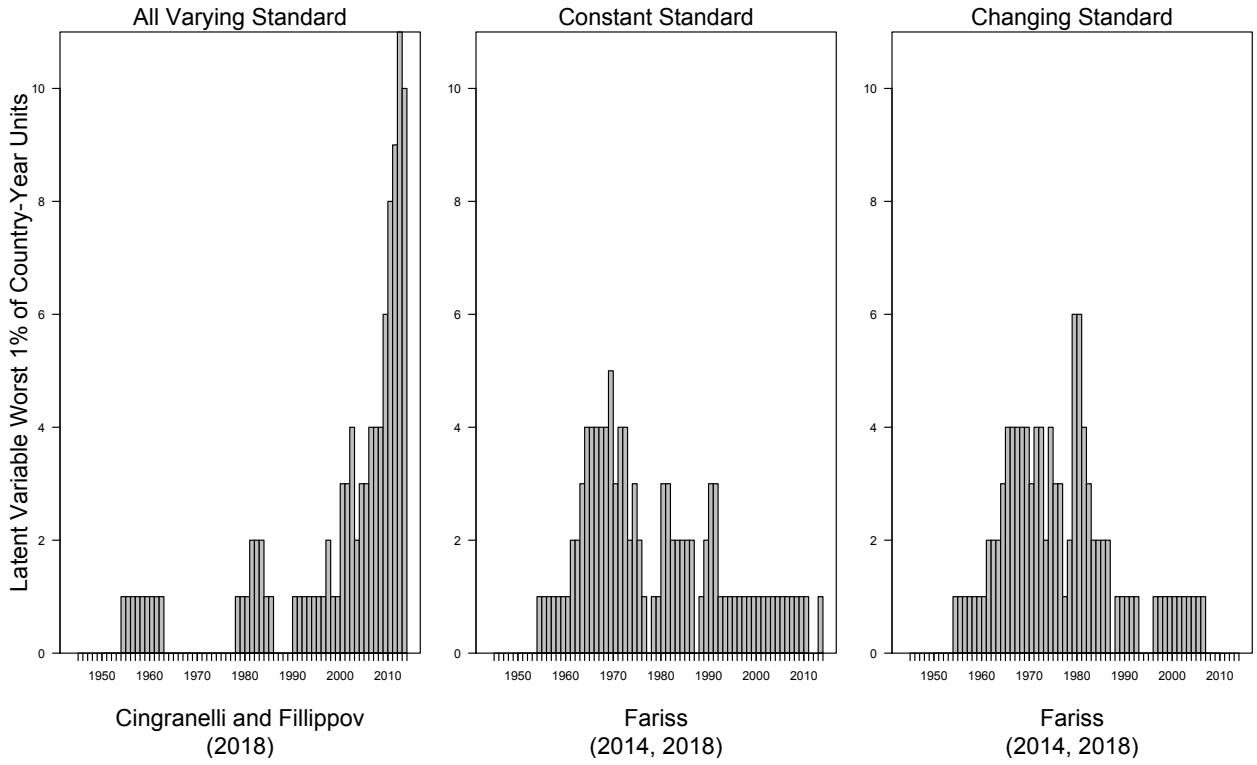

Figure 5: The stacked units in the barplot are the country-years that have the 1% worst scores from three models (all-varying model, constant standard model, changing standard model). The constant standard model and changing model are in close in agreement about which cases are in the top 1% worst. A binary indicator for whether or not the case falls in this worst case category for these two models correlates at about 0.83. The same indicator for the all-varying model correlates with either the constant standard model or changing standard model at about 0.50 or 0.43 respectively. The all-varying standard model proposed by Cingranelli and Filippov (2018) leads to an inference that the worst levels of human rights abuse have just occurred or are possibly even yet to come because it is not identified with respect to time.

4.3 Posterior Predictions of the Yearly-Means for Each of the Observed Human Rights Variables

As I discussed in the simulation analysis section, the yearly means for latent estimates from the all-varying standard model proposed by Cingranelli and Filippov (2018) are reset to 0. This means that any inference about changes in the average levels of the observed human rights indicators is not possible even though we observe such variation over time for each variable (see the appendix for graphs of each variable). Figure 6 displays the differences in correlation coefficients calculated between the yearly mean for each of the 16 observed human rights variables and the estimated yearly mean of one of the three latent variables. The evidence demonstrates the greater explanatory ability of the changing standard of accountability model relative to the all-varying model and constant standard model. The all-varying standard model does poorly because it resets the mean value of the latent estimate to 0 each year making it unable to account for changes over time for any of the variables except for the Rummel events-based variable and also the Harff and Gurr events-based variable.

Spearman Correlation Between Yearly Observed Variable Means and Latent Variable Means

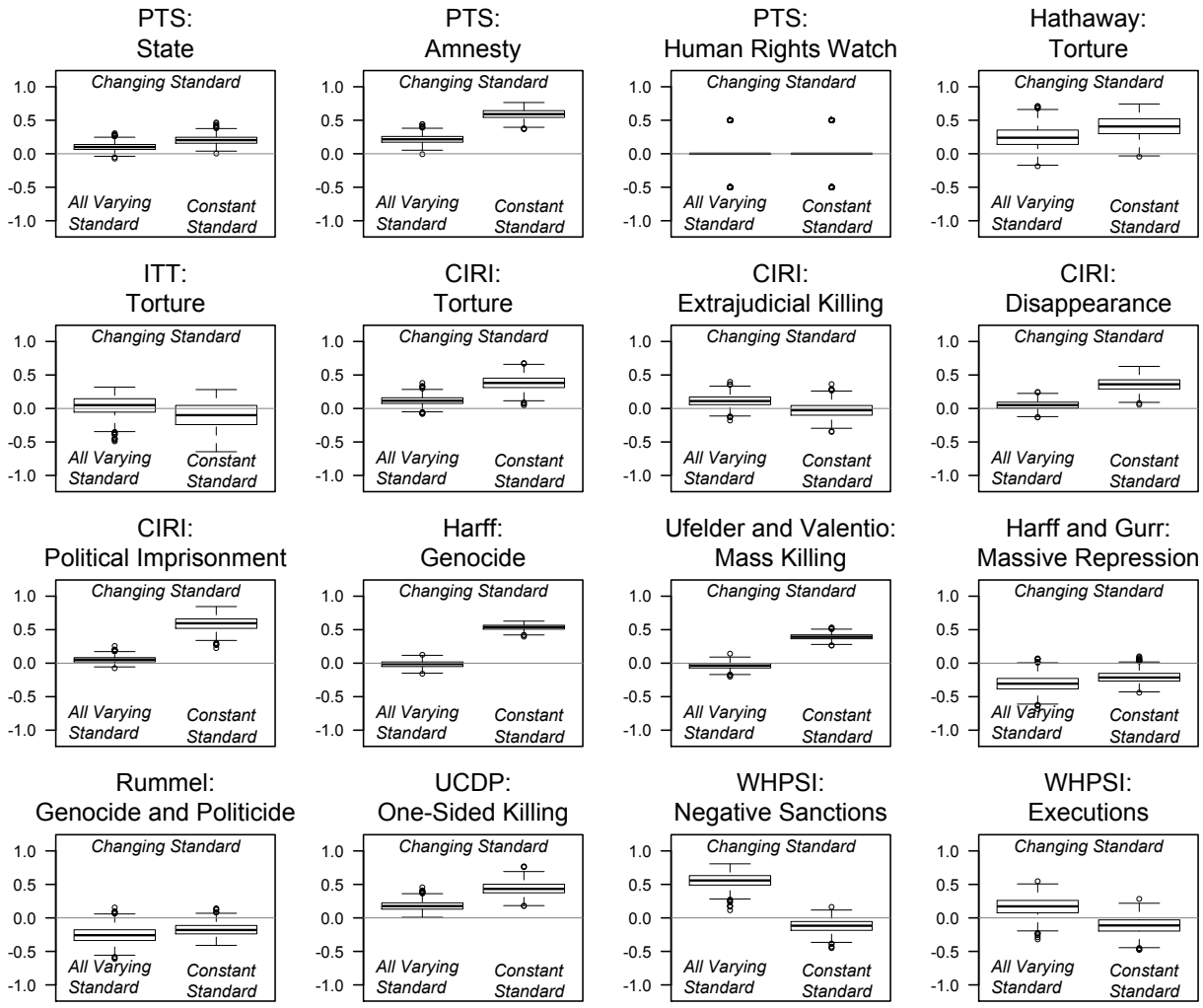

Figure 6: Each panel displays the difference in Spearman correlation coefficients calculated between the yearly mean for the observed human rights variable and the estimated yearly mean of one of the three latent variables. Positive values for either distribution indicate a stronger relationship between the yearly mean for the observed human rights variable and the estimated mean from the changing standard of accountability model. Negative values indicate a stronger relationship for either the all-varying standard model (left distribution) or the constant standard model (right distribution) compared to the changing standard of accountability model. The all-varying standard model does poorly because it resets the mean value of the latent estimate to 0 each year so it is unable to account for changes over time for any of the variables except for the Rummel events-based variable and also the Harff and Gurr events-based variable.

4.4 Alternative Constant Standard Latent Variable Estimates Show Improvements Over Time

Cingranelli and Filipov report that a model with only the standards-based human rights variables shows a stagnant trend in human rights respect over time. This is true. However, not all of the standards-based variables show a stagnant trend in their original categorical form (see the Appendix) or as part of constant standard latent variable models, which I consider in more detail here.

To unpack the difference between the yearly averages for the changing standard model and the constant standard model, I have estimated several alternative versions of the constant standard of accountability model that make use of only the standards-based variables in the Appendix. I begin with a model with just four of these variables and add one additional variable per model. The order in which each new standards-based variable enters the latent variable model are based on evidence from Fariss (2014). Specifically, in Figure 4, Figure 5, and Appendix F of that article, Fariss (2014) shows which human rights variables are the most sensitive to the changing standard of accountability.

Recall that Fariss (2014) shows that the ITT torture and ill treatment variable and the CIRI political imprisonment variable were consistently documented from year to year relative to the event-based variables included in the standard of accountability model. I begin with these variables that are constant over time with respect to the relative frequencies of the event-based variables in addition to the new PTS Human Rights Watch variable, which only spans three years of coverage. I then estimate the constant standard latent variable models (one intercept or one set of cut-points per item), adding in new items in the following order: CIRI Disappearance, PTS Amnesty, CIRI Extrajudicial Killing, Hathaway Torture, CIRI torture, and finally PTS State Department. The variables that are the most sensitive to the changing standard of accountability are the Hathaway torture, the CIRI torture, and the PTS State Department. Only when all of these observed variables are included in the estimation of the yearly average of the latent variable, does the trend line flatten out and become stagnant in Figure 7. This reduction in the slope of the latent variable model as new variables are added is similar to the change when the same standards-based variables are added to the constant standard model using the 7 event-based variables in Figure 8. These models are estimated beginning with the 7 event-based variables displayed in the Appendix. Again, all of these models are estimated with constant item-difficulty cut-points for each versions of the constant

standard model.

The yearly patterns reported here suggest that the Amnesty International reports are more consistently produced from year to year than the State Department reports relative to the event-based repression variables. And, within the State Department reports, that allegations of torture and ill treatment is the topic most sensitive to the affects of the changing standard of accountability (see the Appendix for graphs of the yearly cut-points and probabilities of each category of all of the observed human rights variables). It may be more difficult for monitoring organizations to detect torture and ill treatment in comparison to the other forms of physical integrity rights abuse as the scale of other abuses increases (Brysk, 1994) and easier as the scale decreases (Eck and Fariss, 2018).

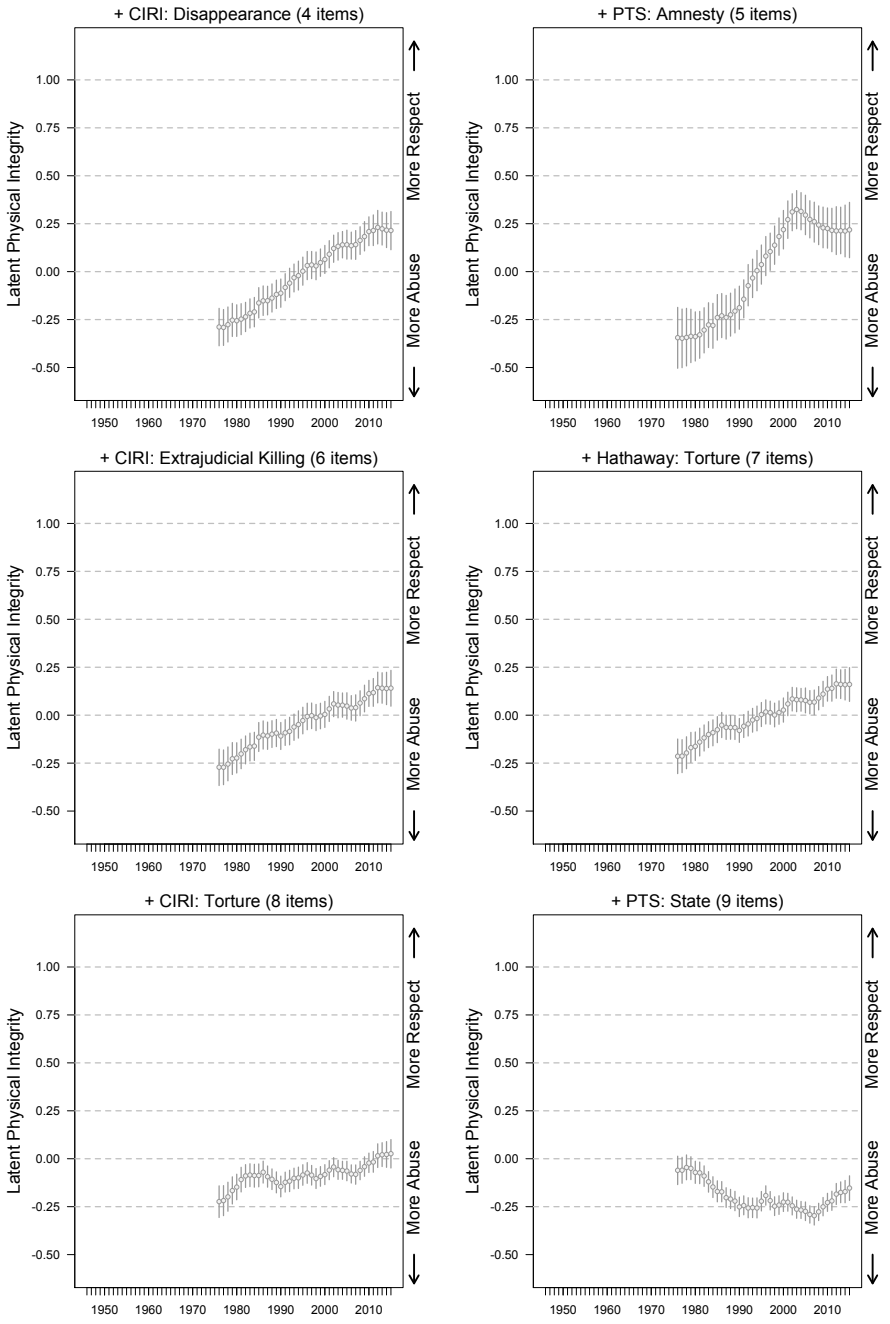

Figure 7: Trends in latent variable estimates for models based on just the standards-based human rights variables over time. All of the models are estimated with constant item difficulty cut-points (constant standard of accountability). The baselines models begin with the PTS HRW, ITT torture, and CIRI political imprisonment. These variables change the least relative to the baseline event-based variables (see Fariss (2014) Appendix F for the statistics that demonstrate the relative strength of the change over time for the standards-based items). Beginning from the upper left panel, one additional standards-based variable is added to the latent variable model in order: 3 + CIRI Disappearance, 4 + PTS Amnesty, 5 + CIRI Extrajudicial Killing, 6 + Hathaway Torture, 8 + CIRI torture, and finally 9 + PTS State Department. Only when all of these observed variables are included in the estimation of the latent variable, does the trend line flatten out and become stagnant.

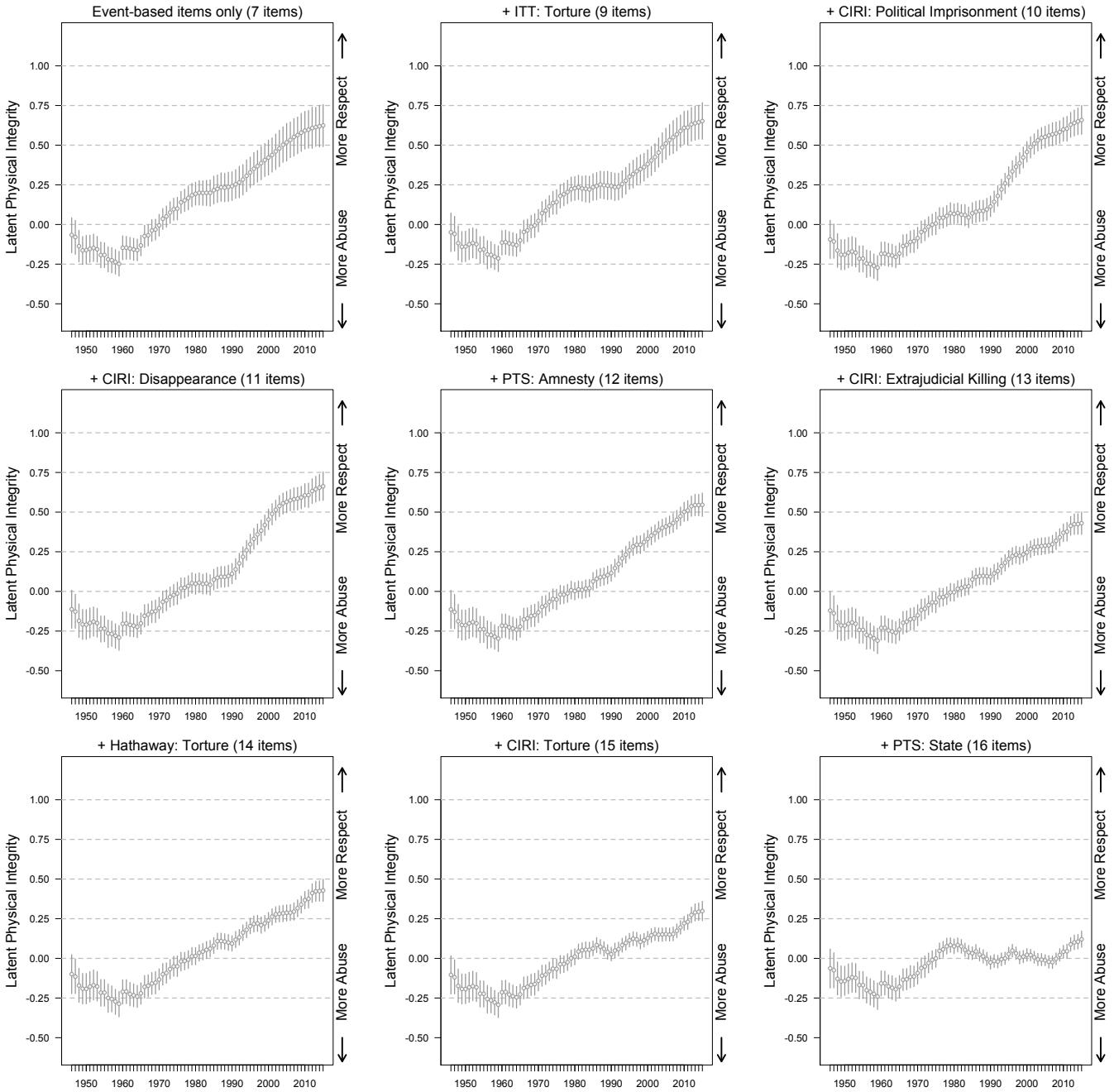

Figure 8: Trends in latent variable estimates for models based on the event-based human rights variables over time with an additional standards-based item. All of the models are estimated with constant item difficulty cut-points (constant standard of accountability). The baselines models begin with the 7 event-based variables and then one additional standards-based variables is added in the same order as above and based on Fariss (2014) (see Appendix F for the statistics that demonstrate the relative strength of the change over time for the standards-based items). Beginning from the upper left panel, one additional standards-based variable is added to the latent variable model in order: 7 items + PTS HRW (not shown for space reasons), 8 items + ITT torture, 9 items + CIRI Political Imprisonment, 10 items + CIRI Disappearance, 11 items + PTS Amnesty, 12 items + CIRI Extrajudicial Killing, 13 items + Hathaway Torture, 14 items + CIRI torture, and finally 15 items + PTS State. As with the latent variable models that only include the standards-based variables, only when all of these observed variables are included in the estimation of the latent variable, does the trend line flatten out and become stagnant.

4.5 VDEM Human Rights Variables Show Improvement Over Time

Finally, in Figure 9, the VDEM torture and killing variables show a substantial increase in respect after the end of the Cold War, which is discussed in another response to Cingranelli and Filippov (2018) (Fariss, 2018*a*). The trend from the VDEM human rights variables are consistent with the latent human right variable that incorporates the changing standard of accountability (Fariss, 2018*a*) and the new trends presented in this section.

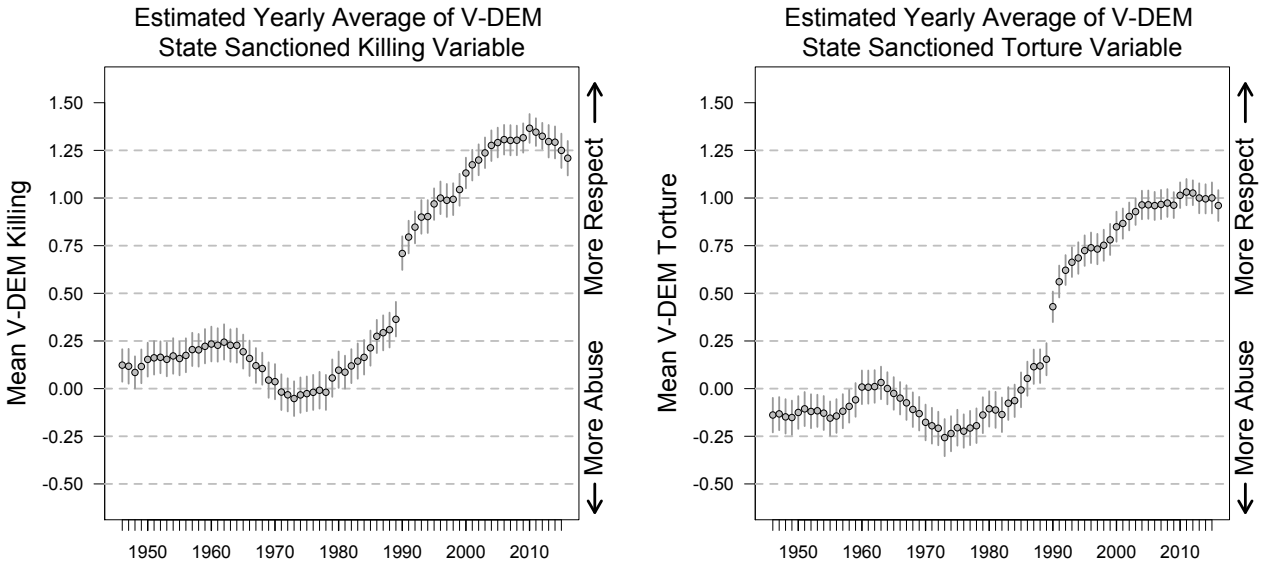

Figure 9: Modified from Fariss (2018*a*): The yearly average for the two expert-coded V-DEM physical integrity variables from 1946-2015 (Coppedge et al., 2014; Pemstein, Tzelgov and Wang, 2015), which is the same time period available for the most recent update of latent human rights variable. What should be clear from this visualization, is a very similar upward trend in human rights respect after the end of Cold War. This upward trend is consistent with the pattern of the latent variable that accounts for the changing standard of accountability first reported in Fariss (2014). These similar patterns provide evidence of the convergent validity of the latent human rights variable that incorporates the changing standard of accountability. According to the V-DEM human rights data and consistent with previous findings (Fariss, 2014), human rights are improving over time.

5 Conclusion

The new human rights estimates developed by Fariss (2014) and extended in this article, supports the conclusion that yes, human rights practices are improving over time. However, these new findings are only possible because of the years of reliable coding conducted by the various coding teams discussed above. Until the publication of the theory of the changing standard of accountability and the new latent variable estimates by Fariss (2014), the academic discourse around human rights progress was becoming increasingly pessimistic (Hopgood, 2013; Moyn, 2010; Posner, 2014). This is because, for the past fifteen years, scholars have puzzled over the stagnating trend in country-year human rights and the negative correlation between UN human rights treaty ratifications and human rights (Hathaway, 2002; Hafner-Burton and Tsutsui, 2005). The categorical indicators from each of these data projects are based on the documentary source material which is changing over time. Thus, the negative patterns are not valid because the data did not account for changes in the source material used to generate the categorical data in the first place (Fariss, 2014, 2018*a*,*b*).

Today, there is reason for new hope, new theorizing, and new data collection. Human rights are getting better and international law seems to be making a difference (Dancy and Fariss, 2017; Fariss and Dancy, 2017) and this result is being corroborated by new and independently generated data from VDEM and by other scholars working on understanding the relationship between law and human rights (e.g., Dancy, 2016, 2017; Sikkink, 2011, 2017). Today, a new and growing literature is working to understand and document how these mechanisms work to change the standard of accountability over time and how these processes manifest themselves as incompletely observable pieces of information encoded in the text of human rights reports (e.g., Bagozzi and Berliner, 2016; Eck and Fariss, 2018; Fariss et al., 2015; Park, Greene and Colaresi, 2017). This is the promise of the science of human rights (Schnakenberg and Fariss, 2014).

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Bagozzi, Benjamin and Daniel Berliner. 2016. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of U.S. State Department Human Rights Reports." *Political Science Research and Methods* DOI: https://doi.org/10.1017/psrm.2016.44.

Brysk, Alison. 1994. "The Politics of Measurement: The Contested Count of the Disappeared in Argentina." *Human Rights Quarterly* 16(4):676–692.

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):407–417.

Cingranelli, David L., David L. Richards and K. Chad Clay. 2015. "The Cingranelli-Richards (CIRI) Human Rights Data Project Coding Manual Version 2014.04.14.".
URL: *http://www.humanrightsdata.com/p/data-documentation.html*

Cingranelli, David L. and Mikhail Filippov. 2018. "Are Human Rights Practices Improving?" *American Political Science Review* doi:10.1017/S0003055418000254.

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):539–568.

Coppedge, Michael, John Gerring, Stafan I. Lindberg, Jan Teorell, Daniel Pemstein, Eitan Tzelgov, Yi ting Wang, Adam Glynn, David Altman, Michael Bernhard, M. Steven Fish, Allen Hicken, Kelly McMann, Pamela Paxton, Megan Reif, Svend-Erik Skaaning and Jeffrey Staton. 2014. "V-Dem: A New Way to Measure Democracy." *Journal of Democracy* 25(3):159–169.

Dancy, Geoff. 2016. "Human rights pragmatism: Belief, inquiry, and action." *European Journal of International Relations* 22(3):512–535.

Dancy, Geoff. 2017. "Deals with the Devil? Conflict Amnesties, Civil War, and Sustainable Peace." *International Organization* TBD.

Dancy, Geoff and Christopher J. Fariss. 2017. "Rescuing Human Rights Law from International Legalism and its Critics." *Human Rights Quarterly* 39(1):1–36.

Dancy, Geoff and Verónica Michel. 2015. "Human Rights Enforcement From Below: Private Actors and Prosecutorial Momentum in Latin America and Europe." *International Studies Quarterly* DOI: 10.1111/isqu.12209.

Davenport, Christian. 2007. *State Repression and the Domestic Democratic Peace*. New York: Cambridge University Press.

Davenport, Christian and Patrick Ball. 2002. "Views to a kill - Exploring the implications of source selection in the case of Guatemalan state terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Eck, Kristine and Christopher J. Fariss. 2018. "Ill Treatment and Torture in Sweden: A Critique of Cross-Case Comparisons." *Human Rights Quarterly* 40.

Eck, Kristine and Lisa Hultman. 2007. "Violence Against Civilians in War." *Journal of Peace Research* 44(2):233–246.

Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability in Human Rights Documents." *American Political Science Review* 108(2):297–318.

Fariss, Christopher J. 2018*a*. "Are Things Really Getting Better?: How To Validate Latent Variable Models of Human Rights." *British Journal of Political Science* 48(1):275–TBD.

Fariss, Christopher J. 2018*b*. "Human Rights Treaty Compliance and the Changing Standard of Accountability." *British Journal of Political Science* 48(1):239–272.

Fariss, Christopher J., Fridolin J. Linder, Zachary M. Jones, Charles D. Crabtree, Megan A. Biek, Ana-Sophia M. Ross, Taranamol Kaur and Michael Tsai. 2015. "Human Rights Texts: Converting Human Rights Primary Source Documents into Data." *PLOS ONE* 10(9):e0138935.

Fariss, Christopher J. and Geoff Dancy. 2017. "Measuring the Impact of Human Rights: Conceptual and Methodological Debates." *Annual of Law and Social Science* 13:273–294.

Fariss, Christopher J. and Keith Schnakenberg. 2014. "Measuring Mutual Dependence Between State Repressive Actions." *Journal of Conflict Resolution* 58(6):1003–1032.

Gibney, Mark, Linda Cornett, Reed Wood, Peter Haschke, Daniel Arnon and Attilio Pisanò. 2017. "The Political Terror Scale 1976-2016." *Political Terror Scale* .

Goldstein, Robert Justin. 1978. *Political Repression in Modern America, From 1870 to Present*. Cambridge, MA: G. K. Hall.

Guitiérrez-Sanín, Francisco and Elisabeth Jean Wood. 2017. "What Should We Mean by 'Pattern of Political Violence'? Repertoire, Targeting, Frequency, and Technique." *Perspectives on Politics* 15(1):20–41.

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):1373–1411.

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):359–371.

Hathaway, Oona A. 2002. "Do human rights treaties make a difference?" *Yale Law Journal* 111(8):1935–2042.

Hopgood, Stephen. 2013. *The Endtimes of Human Rights*. Ithaca, NY: Cornell University Press.

Krüger, Jule, Patrick Ball, Megan E. Price and Amelia Hoover Green. 2013. It Doesn't Add Up: Methodological and Policy Implicaitons of Conflicting Casualty Data. In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, ed. Taylor Seybolt. Oxford University Press.

Mayerfeld, Jamie. 2016. *The Promise of Human Rights: Constitutional Government, Democratic Legitimacy, and International Law*. University of Pennsylvania Press.

McCormick, James M. and Neil J. Mitchell. 1997. "Human rights violations, umbrella concepts, and empirical analysis." *World Politics* 49(4):510–525.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis*. The Hague: Mouton.

Moyn, Samuel. 2010. *The Last Utopia: Human Rights in History*. Cambridge, MA: The Belknap Press of Harvard University Press.

Park, Baekkwan, Kevin Greene and Michael Colaresi. 2017. "The Ups and Downs of Human Rights: Using Aspect-based Sentiment Analysis and Document Meta-data to Explore Information Effects in Human Rights Reports." *working paper* .

Pemstein, Daniel, Eitan Tzelgov and Yi-ting Wang. 2015. Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys. Working Paper 1 The Varieties of Democracy Institute Gothenburg: .

Posner, Eric A. 2014. *The Twilight of Human Rights Law*. Oxford University Press.

Reuning, Kevin, Michael R. Kenwick and Christopher J. Fariss. 2018. "Exploring the Dynamics of Latent Variable Models." *Political Analysis* Conditionally Accepted.

Rummel, Rudolph J. 1994. *Death by Government: Genocide and Mass Murder in the Twentieth Century*. New Brunswick, NJ: Transaction Publishers.

Schnakenberg, Keith E. and Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.

Sikkink, Kathryn. 2011. *The Justice Cascade: How Human Rights Prosecutions Are Changing World Politics*. The Norton Series in World Politics.

Sikkink, Kathryn. 2017. *Evidence of Hope*. Princeton, NJ: Princeton University Press.

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: Yale University Press.

Taylor, Charles Lewis and Michael C. Hudson. 1972. *World Handbook of Political and Social Indicators, Second Edition*. New Haven: Yale University Press.