Line by line response to Cingranelli and Filippov

Christopher J. Fariss*

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

Introduction to the Line by line response

This supplementary document accompanies a new manuscript: "Yes, Human Rights Practices Are Improving Over Time" This paper is an attempt to clarify several points about the theory and model developed in an earlier paper by Fariss (2014). This appendix provides additional discussion in relationship to three critiques: one published and two versions of a working paper, all written by Cingranelli and Filippov. In each case, these authors write in critique of the use of a particular specification of a latent variable model to understand patterns of human rights over time (Fariss, 2014). Unfortunately, the two newer critiques contains many factually incorrect statements. Because of this, I have written a line-by-line rebuttal below.

The original version of the critique is found in section 1 below. The newer version, which makes several new arguments, is available in section 2 below. In section 2, I have noted where my responses correspond to lines from the original critique. Nearly every line from the draft of the critiques is presented below with my responses. The statements from Cingranelli and Filippov appear in italics. My responses are itemized bullet points that follow each of the italicized statements from the Cingranelli and Filippov paper. I provide relevant citations to the political methods literature and also provide visualizations where applicable. Additional evidence can be found in the main response to this critique and an additional response to the first critique (Fariss, 2018*a*,*b*).

1 Line by Line Response to Cingranelli and Filippov version 2 (original version)

1.1

Fariss mistakenly relied on an erroneous functional form to estimate latent scores from observable indicators of human rights that treated some input variables differently from others.

- I have made an argument, supported by a theory, which supports the specification of the latent variable model, which includes the differential item functions used in the changing standard of accountability model compared with the constant standard of accountability model.

- There is no such thing as an "erroneous functional form". The choices of functional form is just that, a choice. Hopefully the choices in how a model is to be specified are supported by a theory. Relatedly, the model also must be identified so that the resulting parameters have meaning relative to one another. Cingranelli and Filippov propose a model that does not meet this criterion because the model is not identified with respect to time. This is a point I return to below and discuss at length in the main response to these authors. In brief, the model that Cingranelli and Filippov propose cannot be used for comparison of parameters from year to year.

- The parameterization of the function for each observed human rights variable in the latent human rights models presented in Fariss (2014) are similar to other models of ordered and binary data. Researchers in American politics are making similar modeling choices (e.g., Caughey and Warshaw, 2016; Hare et al., 2015).

1.2

An unintended consequence of that is that only five out of thirteen input indicators computationally matter for the variation in the latent scores.

- This statement is factually incorrect. Every indicator in the model contributes information to the latent variable as evidenced by the item discrimination parameters reported in the supplementary appendix that accompanies the article by Fariss (2014). In brief, variation in each observed human rights variable corresponds to variation in the value of the latent variable (see Figures 1, 2, 3, and 4 below).

- In the supplementary appendix section F that accompanies the article by Fariss (2014), Figure 6, 7, 8, 9, 10, 11, 12, and 13 display four or more panels that illustrate how the changing standard of accountability is influenced by the standards-based repression variables. The item response curves are based on the value of the item difficulty cut-points and item discrimination parameter estimated for each observed human rights variable. These curves would be flat lines if the standards-based variables did not provide information about the relative placement of one country-year unit to all others conditional on the value of the observed variable.

1.3

The indicators previously used in most human rights studies—the ones based on human rights reports issued by the US Department of State and Amnesty International—do not matter.

- This statement is factually incorrect. Again, every indicator in the model contributes information to the latent variable. Variation in each observed human rights variable corresponds to variation in the value of the latent variable. The standards-based variables are essential for providing information for the estimation of the latent trait.

- Again, see the supplementary appendix section F in (Fariss, 2014), Figure 6, 7, 8, 9, 10, 11, 12, and 13. If this statement were factually correct, then the item response curves displayed in each of these figures would be flat lines. The standards-based variables do provide information about the relative placement of one country-year unit to all others conditional on the value of the observed variable.

1.4

The five indicators that do matter — the ones that computationally determine the variation in Fariss's latent scores, on the other hand, have almost never been used in previous human rights studies.

- This is factually incorrect because all of the indicators have been used by different scholars to study repression.
- Specifically, all of the indicators provide meaningful information to the estimation and placement of the latent trait for each country-year unit. I discuss this point in more detail in the main response to Cingranelli and Fillippov (see also section F in the supplementary appendix that accompanies the article by Fariss (2014)).
- In an earlier reply to Cingranelli and Filippov, Fariss (2018*a*) provides a defense for these variables and the operational protocols designed and updated by the various author teams. And in particular, Fariss (2018*a*) provided justification for the modifications of data sources that comes in the form of count data, which is a critical point in both Cingranelli and Filippov papers. The issue of combining event count data is an issue I am working on but not one that is relevant for this discussion. All event-based variables in the models under discussion are binary indicators that are coded 1 if documentary evidence exists in support of the event occurring and 0 if such documentary evidence does not exist.

1.5

The indicators that matter are the binary (zero or one) records of extreme events such as genocides, mass scale killings of civilians, and political executions by oppressive regimes and foreign powers (such as the list of mass killings compiled by Rummel 1994).

- Again, all of the items that enter the model provide useful information for estimating the relative position of each country-year unit in relation to every other country-year unit. It is a factually incorrect statement to suggest that the variation in the latent trait is determined only by the event-based variables and not the standards-based variables.

1.6

The declining frequency of such extreme events is the true driver of the results Fariss presented as evidence of changing standards in accountability of human rights records.

- This statement is partially true. Variation in the event-based variables do not drive all the results reported in the paper. This variation does lead to the difference in yearly means over time however. The declining frequency of the event-based variables relative to the static frequency of several of the standard-based variables from year to year is the key difference in the data and the one that is supported by the theory of the changing standard of accountability, which motivates the new latent variable model for human rights that incorporates this theoretical concept.

- Fariss (2014) argues that the distinction between the two types of variables — event-based and standard-based — is based on the processes by which the information about human rights abuses is produced. The theory is particularly focused on the information production process that takes place through the publication of the yearly human rights reports (e.g., Fariss and Tyson, 2018). This process occurs before the information is used by political scientists to categorize human rights. The academic coding process is not the theoretical focus in Fariss (2014). For the categorical comparisons of the PTS and CIRI data to be valid, these academic teams rely on the consistent application of the same standard year after year by the human rights monitoring organizations when they produce their reports. If the standards these organizations use to document human rights change over time, as I argue they do, then the temporal pattern in the CIRI and PTS data will be biased over time. I return to this point several times below and in the main response.

1.7

Simulations where conventional human rights indicators are replaced with randomly generated numbers or are totally excluded produce an improving trend in human rights latent scores very similar to the one Fariss presents.

- Cingranelli and Filippov compare the yearly averages from the full changing standard model presented in (Fariss, 2014) to a model that replaces the standards based variables with random categorical values. 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. The standards-based variables provide meaningful information in each model that includes them (see Figure 1, 2, 3, and 4). However, the changing standard of accountability model accounts for the frequency of many of the categorical values from the standards based variables relative to the declining frequency of the events-based variables by allowing the item-discrimination parameters to vary from year to year relative to the fixed value of these parameters for the event-based variables. This modeling choices is based on the theory of the changing standard of accountability and how it influences the production of the human rights reports, which I review in the main response.

- Overall, each of the standards based variables provide meaningful information for the placement of each of the country-year units relative to all the others. As the number of indicators from the model is reduced to include only the event-based variables, the amount of uncertainty for the placement of each unit relative to one another increases because there is less information. So the yearly means from the two models will look similar but the amount of uncertainty associated with these means will increase as the number of indicators decreases.

- The simulations proposed by Cingranelli and Filippov create random ordered categorical variables. Because the data are not generated from any underlying model related to the other data, Cingranelli and Filippov have essentially just added random noise to the model. 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.

- The analysis of the event-based data (with or without additional random data) demonstrates only that the standards based indicators do not have a strong bearing on the global means of respect for human rights across time. The changing standard of accountability model does this by assumption because the theory of the changing standard of accountability suggests these human rights documents are biased over time and therefore the data coded from them are not comparable from year to year. These categorical indicators should therefore have a much more modest impact the yearly average over time. Yet this does not, as Cingranelli and Filippov suggest, indicate that the standards based indicators fail to impact the estimates on the latent variable. On the contrary, the standards-based variables provide significant information about the placement of units with respect to the latent trait within each cross section (year) of the data. This is demonstrated via simulation in the main text (Figures 1 and 2).

- It would be possible to create random data (of any type) that changes the trend of the yearly average in any way. The simulated data could increase the trend, flatten it out, or make it decrease. The simulated data could be combined with the event-based-variables. However, none of these options are what Cingranelli and Filippov describe in their paper.

- I conduct a short simulation study in the main response to Cingranelli and Filippov, which is similar to a much more extensive simulation study presented in Reuning, Kenwick and Fariss (2018).

1.8

We conclude that Fariss's assertions concerning human rights improvement and other findings based on his incorrectly obtained scores are misleading. The pattern of his dynamic latent scores does not reflect changes in physical integrity human rights practices such as torture and political imprisonment, but rather mirrors changes in frequencies of extreme events of mass scale killings and genocides.

- This is a factually incorrect statement. The variation in the latent scores for both the models presented in Fariss (2014) each reflect variation in each of the 16 observed variables now included in the latent variable model of human rights (13 of these variables appeared in the models published in this study). I provide detailed information that expands on this point below and in the main manuscript.

- Figure 1, 2, 3, and 4 visually demonstrate of the latent variable point estimates for each category of all 16 observed variables that enter the models. The pattern of the latent variables scores *does* reflect variation in each of the physical integrity human rights.

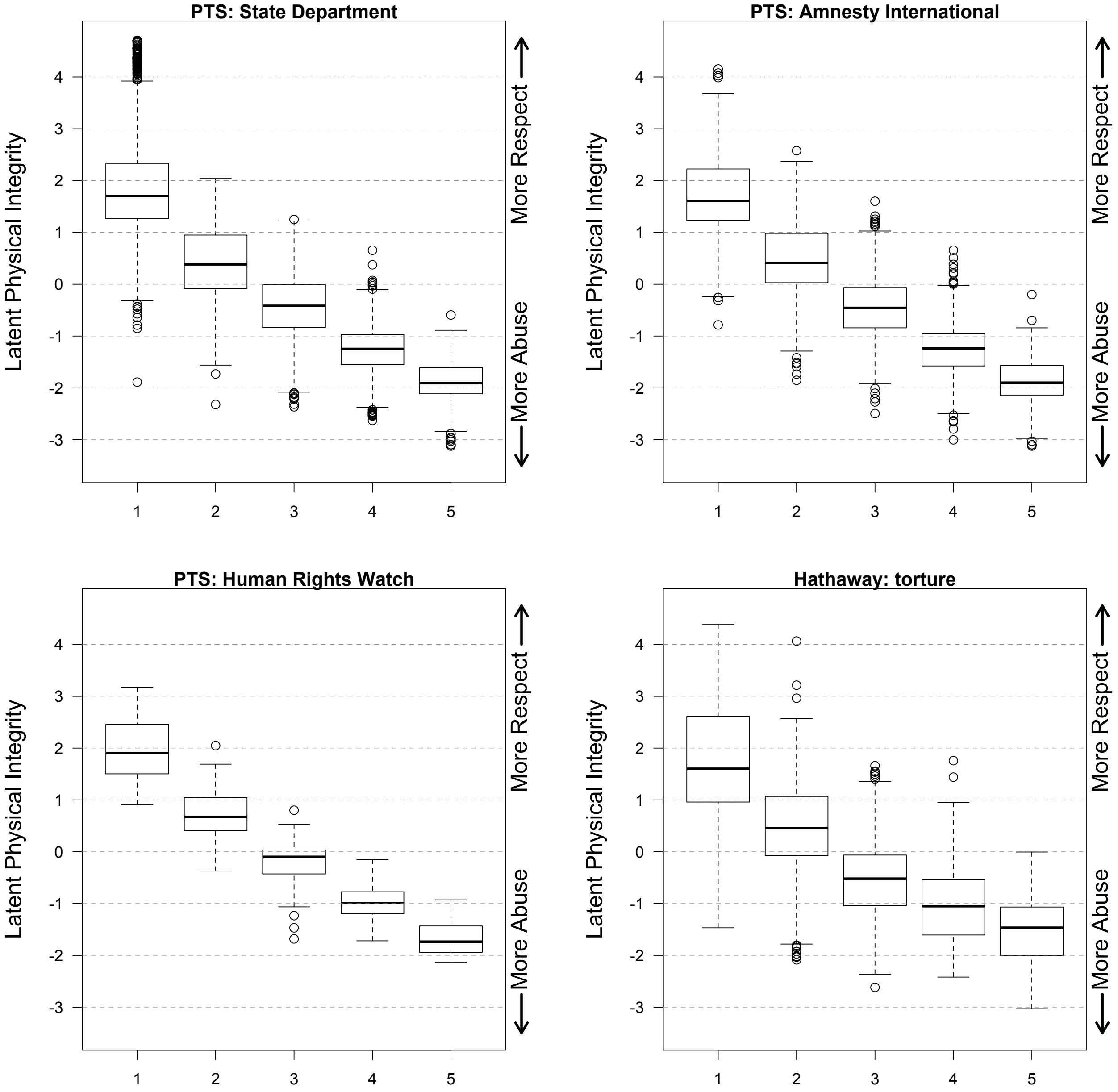

Figure 1: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014) provides detailed documentation for each in the supplementary appendix that accompanies that article.

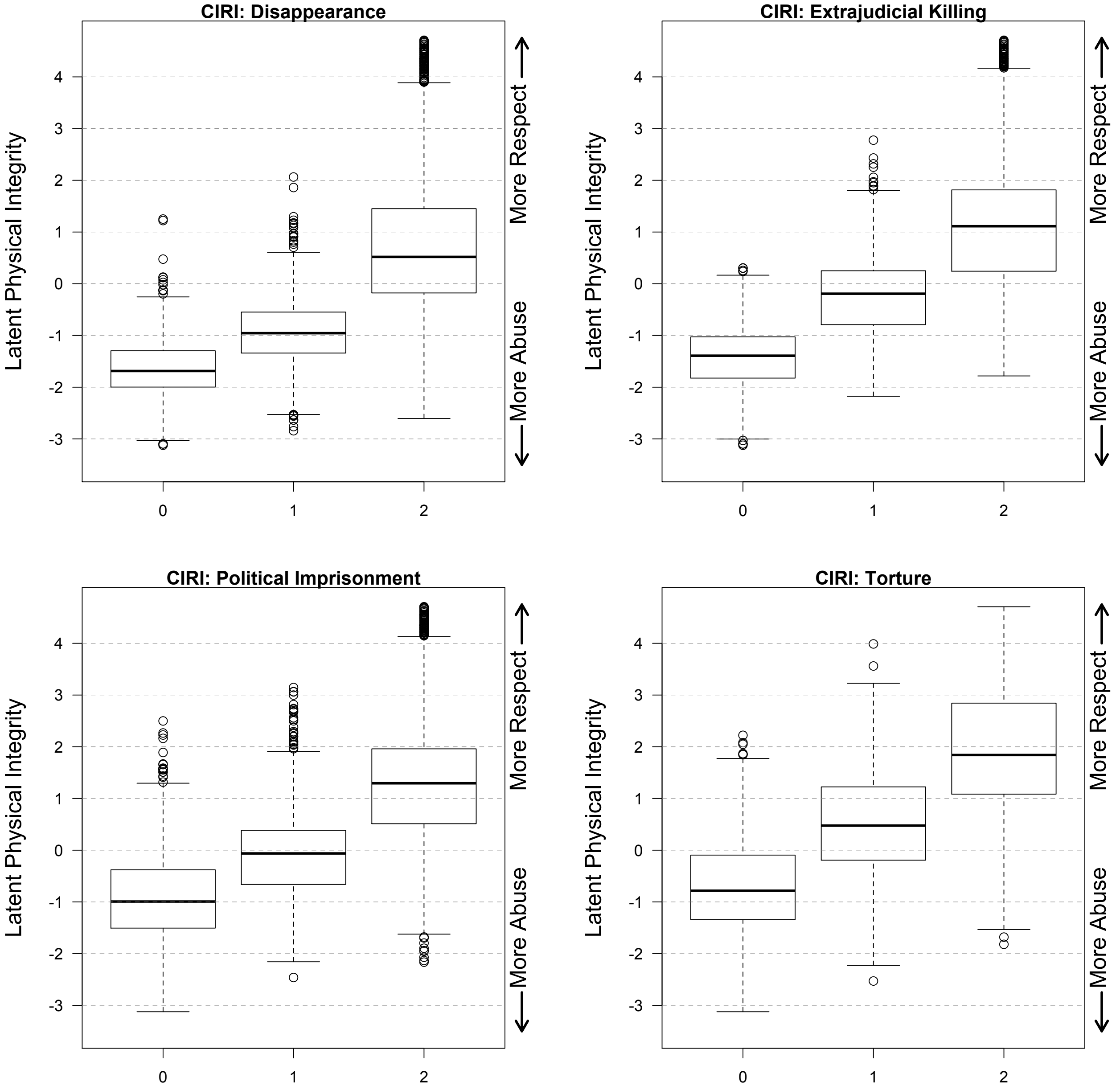

Figure 2: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014) provides detailed documentation for each in the supplementary appendix that accompanies that article.

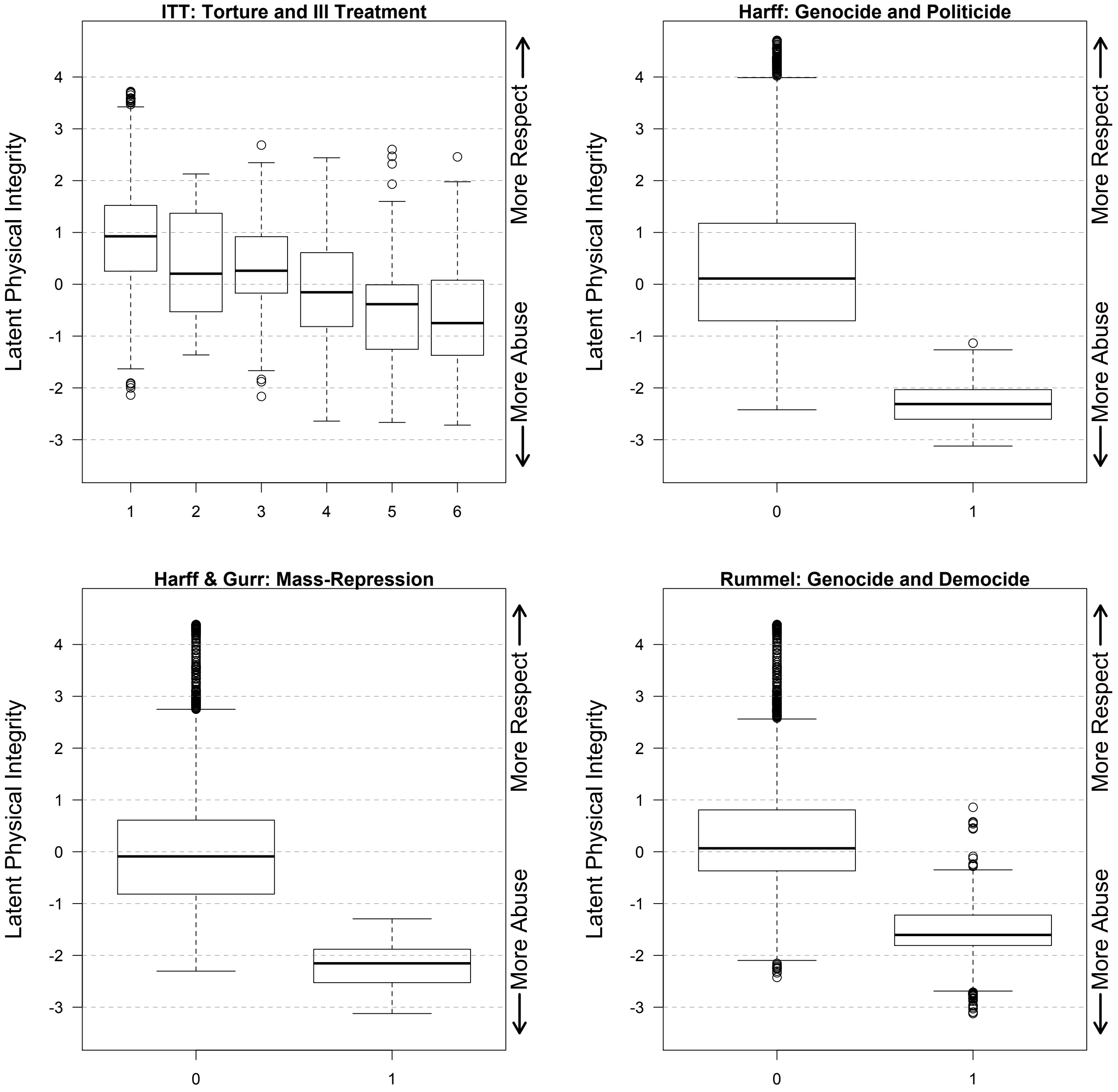

Figure 3: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014) reflects variation in each of the 16 observed variables now included in the model.Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014) provides detailed documentation for each in the supplementary appendix that accompanies that article.

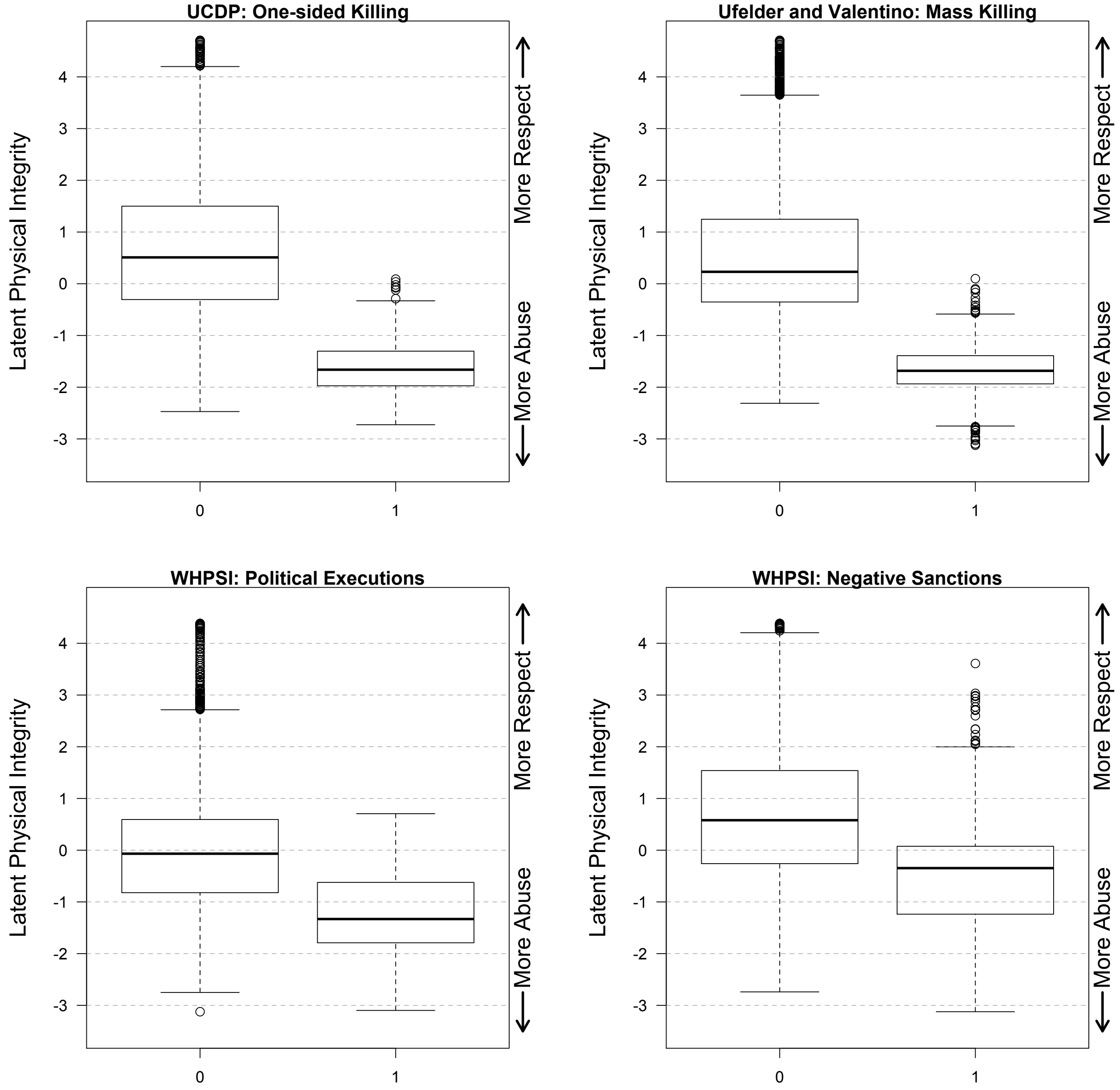

Figure 4: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014) provides detailed documentation for each in the supplementary appendix that accompanies that article.

1.9

First, Fariss wrongly relied on a never before used modification of dynamic IRT where some input variables were treated differently from others.

- This is a factually incorrect statement. The latent variable model is supported by a theory about how some categorical human rights information is produced (the standards-based data) compared to other categorical human rights information (the event-based data). Many applications of latent variable measurement models in political science use theory to support the parameterization of the unobserved theoretical concept (e.g., Clinton, Jackman and Rivers, 2004; Martin and Quinn, 2002; Poole and Rosenthal, 1991, 1997; Poole, 2005).

- For similar models to the one developed in Fariss (2014), interested readers should consult papers that also grapple with similar issues using differential item functions (e.g., Caughey and Warshaw, 2016; Hare et al., 2015), and the hierarchal modeling text by Gelman and Hill (2007) for many additional examples.

1.10

His modification of dynamic IRT makes it is impossible to falsify the claims (hypotheses) he seeks to evaluate, because the erroneous statistical specification produces a predictable data-invariate trend.

- This is another factually incorrect statement. It is possible to falsify the empirical claims made in Fariss (2014).

- The model specification was not constructed in error. It is supported by a rigorous theory and empirical evidence that supports the differentiation of the events-based and standards-based data because of the differences in the process by which this information is generated and used to code the categorical human rights variables. This is a point I discuss at length in the main response to the critique by Cingranelli and Filippov.

- The latent estimates are not invariant to the inclusion of the standards-based data. I discuss the variation that these observed variable provide for the placement of the country-year units relative to one another (see also the supplementary appendix section F in Fariss (2014)).

- Fariss (2014) discusses at length empirical information that demonstrate that not all of the observed human rights variables that are coded from the standardized human rights reports are influenced by the changing standard of accountability. I have reproduced that passage here:

"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 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 Appendix 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 period of coverage (1995-2005) or differences between Amnesty's Urgent Action Reports, which these data are based upon, 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 [changing standard of accountability 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."

1.11

Second, unlike Fariss, previous uses of dynamic IRT in the literature treated all input variables equally. We show that if all available records of human rights are treated equally, the estimated dynamic latent *scores follow patterns similar to the patterns previously reported in the literature.*

- This is true, which is demonstrated with the constant standard model presented in Fariss (2014). I have reproduced this model in the main response along with the changing standard of accountability model and the model proposed by Cingranelli and Filippov (the all-varying standard model). The constant standard model, when all cut-points and intercepts are fixed, is the model that does exactly this. The model proposed by Cingranelli and Filippov is not identified with respect to time. Thus, the comparisons of the latent variable from this model are not comparable year to year, which I demonstrate with a simple simulation analysis in the main response to Cingranelli and Filippov.

- The model proposed by Cingranelli and Filippov is not identified. What is happening with the all varying intercept (cut-point) model proposed by Cingranelli and Filippov, is the mean of the latent trait is being reset to 0 every year. This is the case no matter what the values of the observed human rights variables are. So as conditions get better from one year to the next (or stay unchanged), the model is forced to push the bad cases further down into the negative portion of the latent space so that they are far enough away from the mediocre and good cases, conditional on the available data. This is because the mean for each year must be 0 by assumption in the Cingranelli and Filippov model. The prior position of each unit is informative *within* each year but the model still centers the distribution for of all units in a given year over 0. This is because Cingranelli and Filippov have estimated a model that estimates the intercepts (cut-points) for each item for each year. There is essentially no baseline from year to year so the model has to revert to the mean 0 assumption of the prior distribution for the latent trait. 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 so that the mean estimate for the units in that year is 0.

- Put another way, the model proposed by Cingranelli and Filippov is similar to modeling each year of data as a separate model and then combining the estimates back together. Instead of letting the latent variable model use the distribution to arrange all of the country-year units relative to one another for all years, the all-varying cut-point model is reseting the distribution each year and only arranging the country-year units relative to one another within one year. This is why the estimates are not useful for making comparisons from one year to any other year thought still are suitable for making comparisons within a given year.

1.12

Third, we disagree with Fariss and argue that there are no methodological reasons to privilege records of mass scale repression such as genocide, politicide, or other form of mass killing as more valid and reliable than more conventional indicators of human rights.

- This is not a methodological issue but rather a theoretical one. The standard applied to the production of the human rights by Amnesty International and the State Department are not influenced by political scientists. All that we as political scientists are able to do is directly code the existing documents (standards-based data) or take information from a set of documents to determine whether or not a specific type of repressive event has occurred (event-based data).

- 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 coding. For the standards-based variables, the documents are directly coded. For the event-based variables, documentary evidence is taking from multiple sources and used to look for evidence that a particular type of repressive event occurred. These are fundamentally distinct categorization processes.

- Cingranelli and Filippov have not provided a theoretical argument that supports their claim that the standards-based and events-based variables are produced in the same way. If the source material used to code the event-based variables and the standards-based variables are indeed produced in the same way, then Cingranelli and Filippov need to argue in favor of the constant standard model from Fariss (2014). The model proposed by Cingranelli and Filippov is no identified with respect to time. Only cross-sectional comparisons are possible in the model they propose. I provide more details on this below.

1.13

Finally, many of the new scores, as generated by Fariss, fail face validity, especially those produced through backward data extrapolation and imputation for the 1949-1975 period.

- This is actually an issue of concurrent validity. Face validity is an assessment of the operational protocol itself: will the instrument, test, or the specific questions on the test, be effective at eliciting information from the underlying trait of interest?[1] Another way of thinking about face validity is as a validation technique that links a theoretical concept to the operational protocol used to generate empirical content about that concept. No empirical content is necessary though for this assessment because it is about how closely we believe the operational protocol maps on to the concept embedded in the theory. So, based on my description of the model proposed by Cingranelli and Filippov, which I discussed above, I can say that it is invalid on its face because the model forces the mean of the latent trait in each year to be 0. This is not a good attribute for the operational protocol (the latent variable model is the final part in the operationalization process). The constant standard and changing standard models presented in Fariss (2014) do not force the yearly mean to be 0 (only the global mean for all the country-year units). These two models, as they relate to the theory, therefore have face validity, at least based on this model specific criterion: temporal comparison.

- Concurrent validity on the other hand is an empirical assessment that links the data obtained from the operational protocol to previously obtained or known estimates of the same concept (Adcock and Collier, 2001; Trochim and Donnelly, 2008). Usually though in practice, we use concurrent validity with pre-existing categorical information or rank order data. I know that Sweden should have a stellar human rights record but it is coded as torturing by the CIRI data (Eck and Fariss, 2018). This is a concurrent validity issue that reveals a deviant case, both of which are topics that I discuss more below (see also Fariss (2018*a*)).

- The latent variable model is not an extrapolation. An extrapolation is a prediction of a value for a

[1] Adcock and Collier (2001) prefer to not use the term "face validity" because the definition varies from user to user. Instead, they prefer the term content validity. Content validity is simply a check of the operationalization against the relevant content domain for the theory" (Trochim and Donnelly, 2008).

variable based on data available for the same or similar units in a period of time prior to or after the unit for which the predicted value is required. An interpolation is similar in that observed data for the same or similar unit are available both before and after the unit for which the predicted value is required. In the latent variable model, every unit has observed data which is used to estimate the value of the latent trait. It is a parameter based on the observable data that is available for a given country-year unit. Specifically, each country-year-unit included in the model has at least one observed variable available for it. Every unit is therefore based on observable data. The estimate for each unit, the position it is placed on along the latent trait, is selected with respect to the values of the available observed variables, based on units with other similar values. Since all of the variables relate to the underlying theoretical concept of repression, country-year units, tend to be grouped with states that have similar values on each of the items. When fewer items are available such as the 1946-1975 period, there is greater uncertainty about where to place these units. This uncertainty is captured by the standard deviation for each of the latent estimates. I provide a graph that visualizes this below in Figure 15.

- The latent variable model simply places each of the country-year units relative to one another along a single interval-level dimension. Along the latent trait, a score of 0 means that the particular units that receive this score (or close to it) are average relative to one another based on the available information for those particular country-year units. A unit standard deviations above or below this 0 value has a similar meaning in that the units are more or less distinct from the average unit. These relative placements along the interval level latent trait correspond to values of the items used to estimate these positions. When less information is available, the precision of the placement decreases. That is why it is of paramount importance to acknowledge and include the level of uncertainty for each unit in any analysis. This is a point that is emphasized and described in detail in several published articles (e.g., Fariss, 2014; Schnakenberg and Fariss, 2014).

1.14

In particular, many of Fariss's scores measuring protection of physical integrity rights place authoritarian, less developed countries above the well-established wealthy democracies.

- Fariss (2018*a*) comments on this point at great length in a published response to another critique by Cingranelli and Filippov. In particular, Fariss (2018*a*) discusses the deviant case of the United States in 1953.

- In brief, such a misplacement of the unit on the estimated value of the theoretical concept, is called a deviant case. "A deviant case is an observation that is coded at a surprising value or outlier along some theoretical concept (Lijphart, 1971; Seawright and Gerring, 2008). The identification of such cases does not undercut the progress already made in enhancing the validity of recent versions of the latent human rights variable because each new model has been able to distinguish between theoretically distinct cases that earlier variables were not able to identify" (Fariss, 2018*a*).

- For example, Eck and Fariss (2018) highlight a deviant case in the CIRI torture data: Sweden. Sweden is categorized as a torturing country in 1/3 of the country-year units in the CIRI dataset. It is therefore grouped with countries receiving the same score such as Haiti, Belarus, or Bangladesh. As argued in Fariss (2018*a*), "First, latent variables allow for the exploration of deviant or unexpected cases (e.g., the CIRI human rights data categorizes Sweden in 2011 and Guatemala in 1983 as both engaging in the same level of torture). This type of case study is a productive research design strategy for identifying new theoretical concepts that relate to other sources of bias in the human rights documentary sources. To enhance validity, these theoretical concepts, like the changing standard of accountability, should be incorporated into future versions of the latent human rights model."

- I have incorporated several additional indicators in the latent variable model of human rights. The inclusion of these new variables addresses the issue raised by Cingranelli and Filippov. Two of these new indicators are binary event-based variables. With the updated model, the United States in 1953 is still a case with serious human rights issues, but it's placement on the updated latent variable is not as surprising as it once was.[2] These case study designs, which provide a type of concurrent validity assessment, are useful for learning how well improvements to the latent variable model work. Below I will discuss such issues with respect to the alternative model specification

[2] Eck and Fariss (2018) also discuss cross sectional comparability and organizational differences between the monitoring organizations themselves.

proposed by Cingranelli and Filippov.

1.15

Cingranelli and Filippov mischaracterize the theory from Fariss (2014) stating:

Human 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.

- This interpretation of the argument from Fariss (2014) is more consistent with the argument made by Clark and Sikkink (2013).
- As Clark and Sikkink (2013) argue, coders may be influenced by larger quantities and greater quality of information when coding human rights reports. The argument made by Fariss (2014) builds on this idea but shifts the conceptual focus away from the coders and towards the producers of the human rights reports themselves: The US State Department and Amnesty International. Thus, Cingranelli and Filippov misunderstand the theory of the changing standard of accountability. It is not about academic coding procedures. Instead, it is about the monitoring agencies and the way the human rights reports themselves are produced. A coding procedure can be applied with complete fidelity and the changing standard of accountability can still influence what CIRI and PTS scores a particular case receives. The standards-based data are potentially biased not because the coding procedure is biased but because the reports themselves are produced by monitoring agencies that are changing the standards that they use in the process of documenting human rights abuse.

1.16

Dynamic versions of IRT assume that criteria for recording the indicators could change over time.

- This is not an assumption of the model. In fact the opposite assumption is true. The latent variable model, with fixed item difficulty parameters (intercepts or cut-points) assumes that the observed data are generated in the same way for each unit. This is a conditional independence assumption and it means that only variation in the latent trait should be related to the value that the observed variable takes. To use an analogy from the educational testing field, the test should be the same difficulty level for every student. Students with greater ability will do better on the test (be placed relatively higher) than students with lesser ability.

- The dynamic prior on the latent variable itself is an assumption about the relationship between units, conditional on the values of the observable items. This is a modeling choice that relaxes another conditional independence assumption which Schnakenberg and Fariss (2014) discuss at length.

- The dynamic latent variable model that incorporates the changing standard of accountability needs an additional assumption about the probability of observing one type of item (standards-based data) *relative* to the probability of observing another type of item (event-based data) (Fariss, 2014). Without this assumption and the evidence that supports making the distinction between the standards-based data and the events-based data, running such a model would not be possible because it is not identified with respect to time. That is, the estimates of the latent traits would not be comparable from year to year if all the intercepts (cut-points) varied as proposed by Cingranelli and Filippov.

1.17

Fariss (2014) introduced his own unique version of dynamic IRT and estimated what he called a dynamic latent index of human rights'.

- The latent human rights variable is not an index. This is a misuse of the term index (see Fariss and Dancy, 2017, for a discussion). All of the ordered variables (including the CIRI variables) that enter the model are scales.

- Again, for similar models to the one developed in Fariss (2014), interested readers should consult papers that also grapple with similar issues using differential item functions (e.g., Caughey and Warshaw, 2016; Hare et al., 2015), and the hierarchal modeling text by Gelman and Hill (2007) for many additional examples.

1.18

This new version of the dynamic IRT method allowed Fariss to demonstrate a new pattern in human rights scores showing that human rights have improved significantly since the 1980s. His estimates show that there was significant reduction in the use of torture over the past decades

- The model does not show that torture specifically has reduced. It shows that the probability of a country-year unit having sufficient documentary evidence available to register as engaging in torture on the CIRI variable increases over time. It also shows that on average, the value of the latent variable improves from year to year.

- The results from the changing standard model also suggests that the instances of ill treatment and torture were likely not catalogued in earlier periods because the monitoring organizations did not have the resources of capacity to record such events. This part of my argument is consistent with an argument and detailed case study evidence from Argentina presented by Brysk (1994) in which the author states: "Incidents 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" (681).

- The results from the changing standard model also suggests that new types of abuses that were not previously classified as ill treatment or torture are also being classified and then catalogued by the monitoring organizations. This is consistent with patterns of topical attention described by recent textual analyses of the human rights reports (e.g., Bagozzi and Berliner, 2016; Fariss et al., 2015; Park, Greene and Colaresi, 2017).

1.19

In contrast, trends produced using CIRI or PTS scores show that the trend in human rights is flat and that more countries are using torture in more recent years (Cingranelli and Richards 2010; Wood and Gibney 2010).

- This comment is misleading and not consistent with the CIRI, PTS, or other human rights data. In Figure 5 and Figure 6, only the CIRI torture variable shows any evidence of a decline. But this decline occurs in the late 1980s and early 1990s. If this is what Cingranelli and Filippov mean when they say that "more countries are using torture in more recent years", then these authors need to state when these periods start and stop more precisely. Specifically, the trend for CIRI torture variable flattens out from about 1995 through 2011, the last year for which the CIRI data are available. This trend is not consistent with other variables measuring torture. The ITT torture variable shows a clear improvement (more respect) in the average level of this variable. The Hathaway torture variable shows a slight decline (less respect), which is consistent with years in which it and the ITT torture variable are both available (1995-1999). In Figure 8, the VDEM torture variable shows a substantial increase in respect after the end of the Cold War, which is discussed in another response to Cingranelli and Filippov (Fariss, 2018*a*). The VDEM human rights variables are consistent with the latent human right variable that incorporates the changing standard of accountability (Fariss, 2018*a*).

- In Figure 5 and Figure 6 none of the other standards-based human rights variables show evidence of a decline. In fact several show evidence, in some cases substantial, of an improvement. Some of these improvements are quite recent. The trend for the CIRI killing variable is flat for the entire time period of the dataset (1981-2011) but both the political imprisonment variable and disappearance variable show evidence of a positive improvement.

- None of this evidence is commensurate with the claim that "In contrast, trends produced using CIRI or PTS scores show that the trend in human rights is flat and that more countries are using torture in more recent years". Moreover, independent data collection from the Varieties of Democracy project corroborate the evidence for improvements for physical integrity rights. These patterns are all consistent with the declining trend in the occurrence of seven event-based human rights variables in Figure 7.

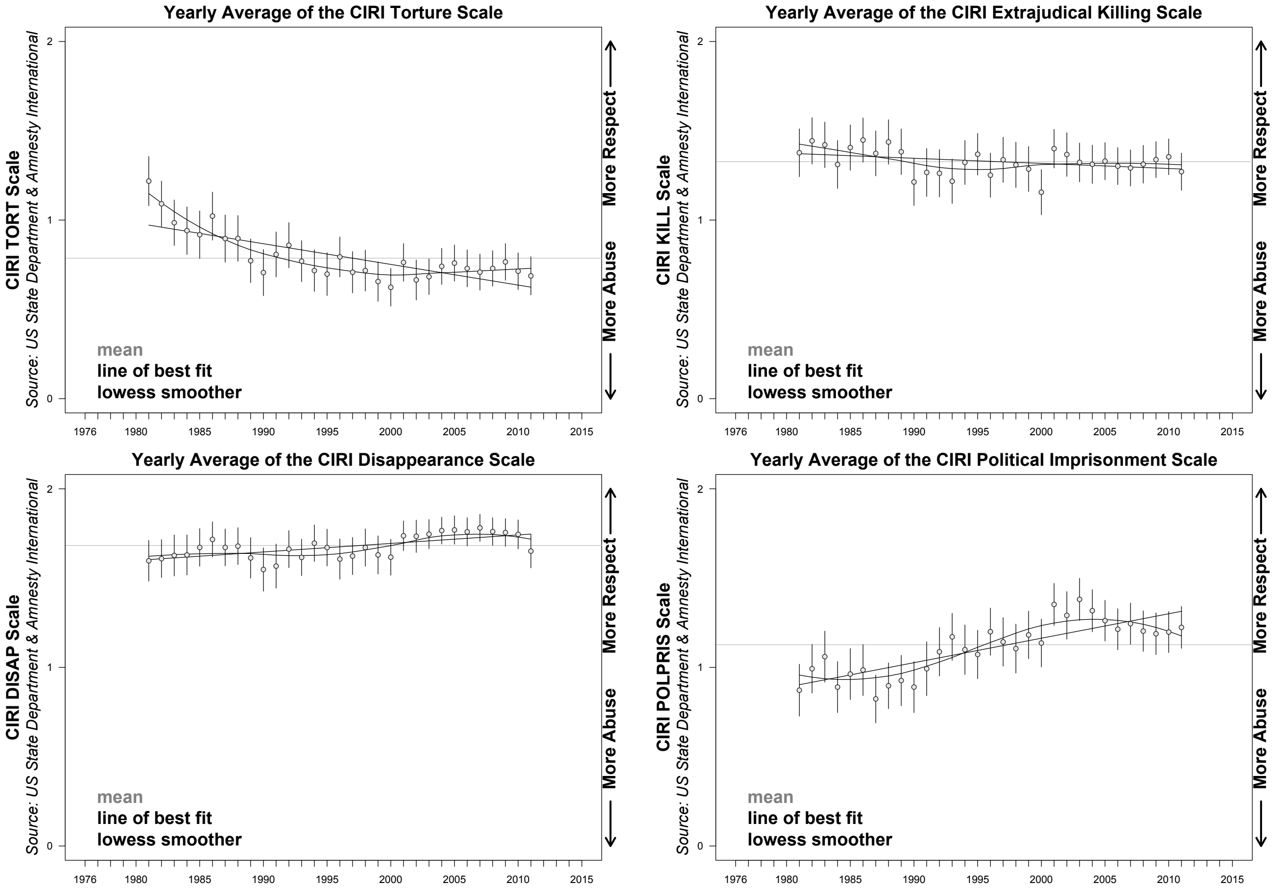

Figure 5: Trends in standards-based human rights variables over time. Note that all of these variables except the ITT and CIRI political imprisonment variables enter the changing standard of accountability model with varying item difficulty parameters (one for each year). The ITT and CIRI political imprisonment variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

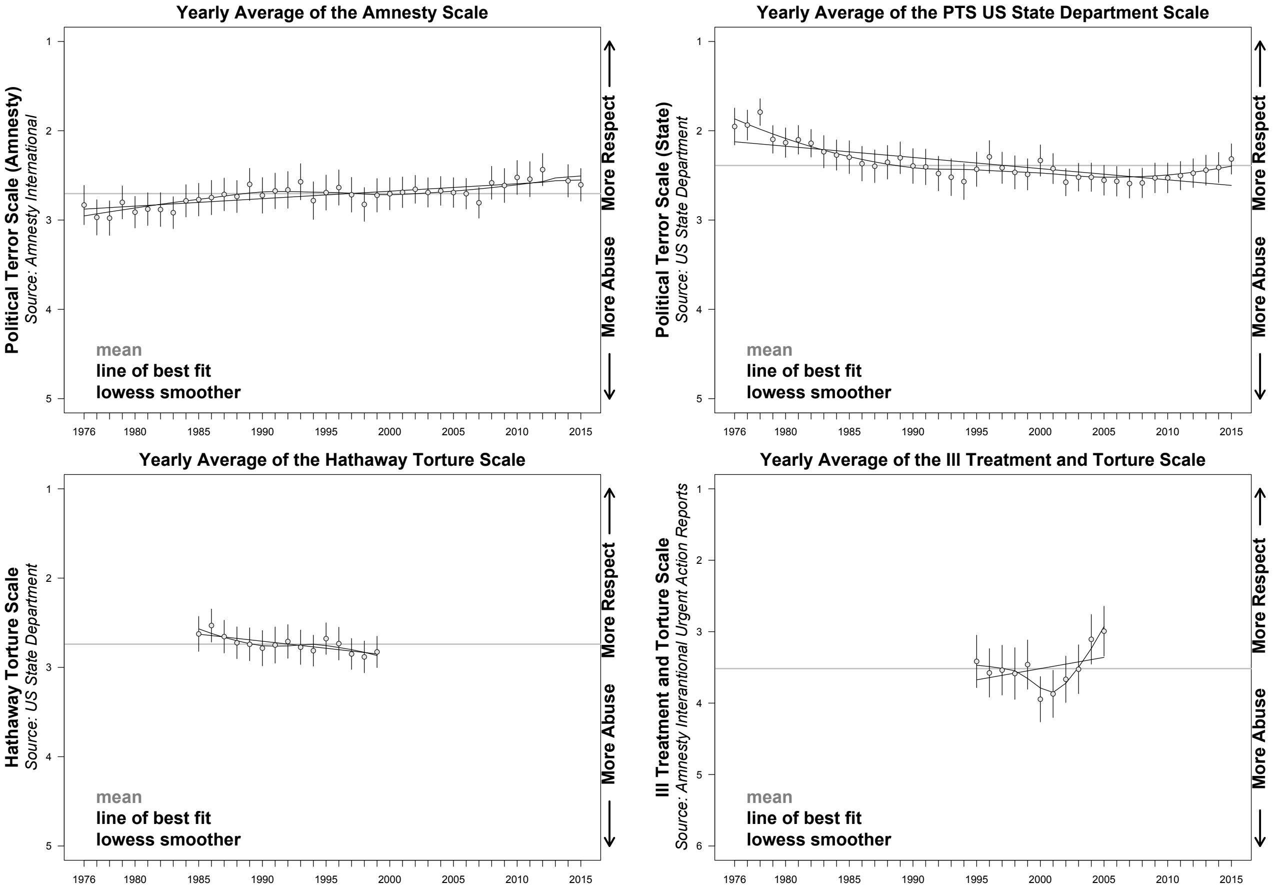

Figure 6: Trends in standards-based human rights variables over time. Note that all of these variables except the ITT and CIRI political imprisonment variables enter the changing standard of accountability model with varying item difficulty parameters (one for each year). The ITT and CIRI political imprisonment variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

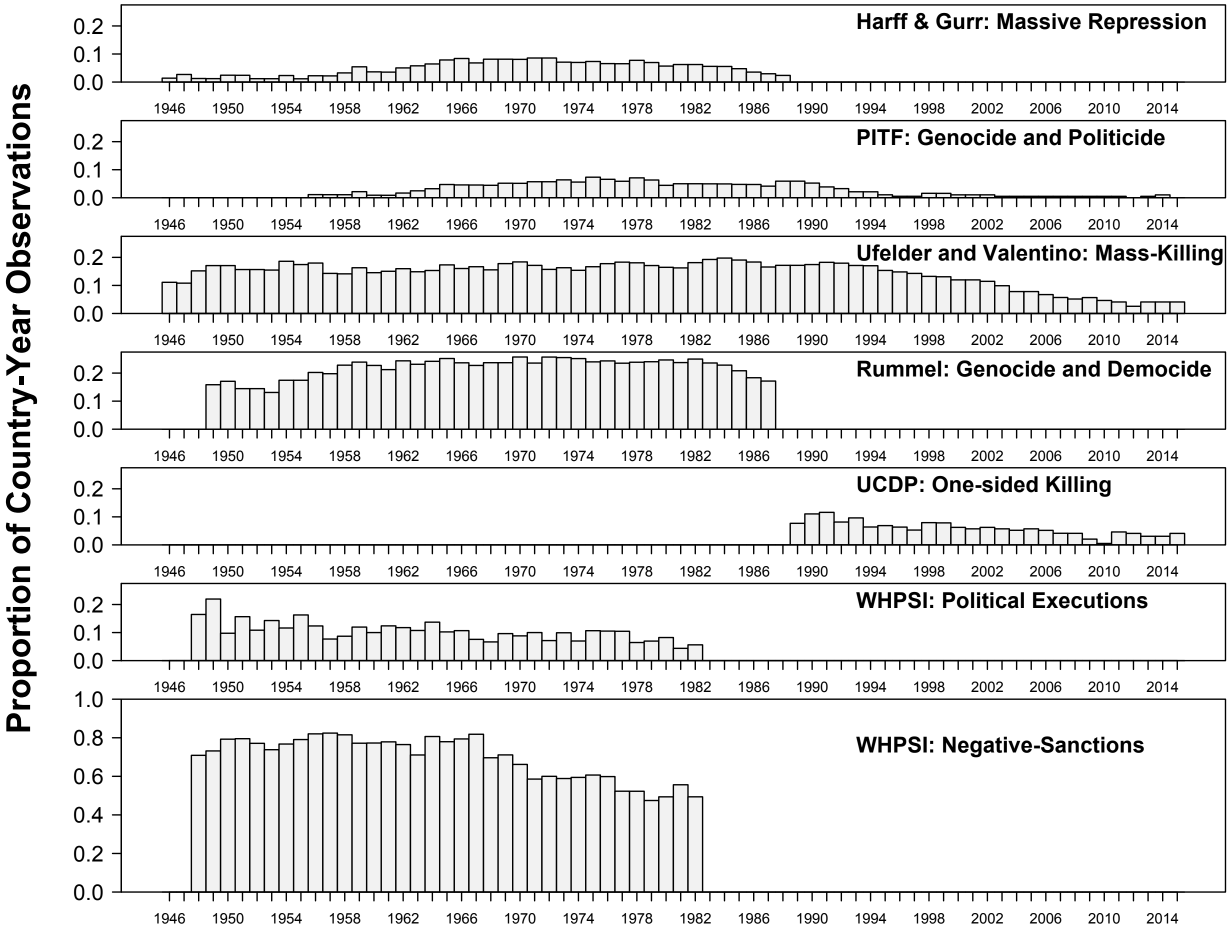

Figure 7: Trends in events-based human rights variables over time. Note that these variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

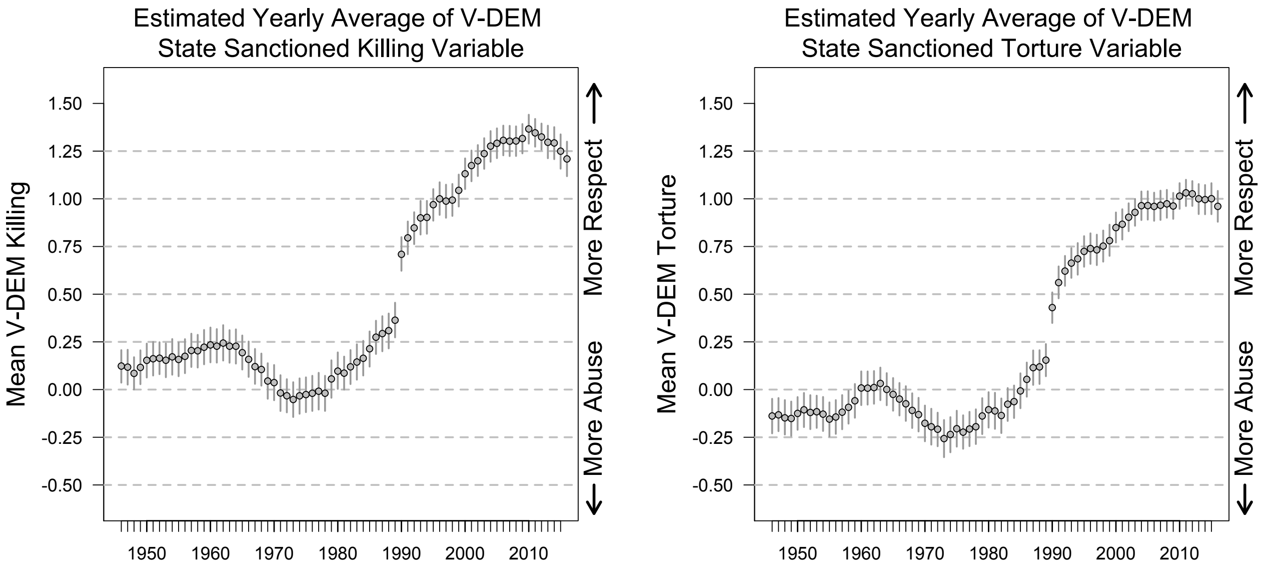

Figure 8: Modified from Fariss (2018*a*): The yearly average for the two expert-coded V-DEM physical integrity variables from 1949-2013 (Coppedge et al., 2014; Pemstein, Tzelgov and Wang, 2015), which is the same time period available for the most recent update of latent human rights variable. What should 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.

1.20

Unfortunately, his erroneous statistical model specification produces a predictable and data-invariate trend.

- The first part this statement is misleading. I did not make an error in the model specification. The model specification was purposely created. Again, it its supported by a rigorous theory and empirical evidence. The empirical evidence is obtained by comparing the model estimates from one model to another using a variety of statistical tools, which are discussed in detail in Fariss (2014), Fariss (2018*b*), Fariss (2018*a*), Schnakenberg and Fariss (2014), and this most recent response.

- In the response to Cingranelli and Filippov, I provide evidence from posterior predictive checks which demonstrates how estimates from the model proposed by Cingranelli and Filippov compare to the constant standard model and the changing standard of accountability model. In summary, their model looks very similar to the constant standard model only because there is no change over time for this model. However, I as I have already explained, the lack of temporal change over time for the model proposed by Cingranelli and Filippov will happen whether or not the underlying data change over time or not.

- The second part of this statement is demonstrably false. The model is not "data-invariate", which is a term that the authors use several times but never define. I presume that they mean that the latent variable estimates are not based on variation in the standards-based variable. This is again false as demonstrated visually in Figures 1, 2, 3, and 4.

- Moreover, the item discrimination parameters presented in the supplementary appendix document (section F), which accompanies the main article from Fariss (2014) provides evidence that each item contributes meaningful information to the placement of each unit along the internal level distribution of the latent variable. The positive values for each item indicates that each of them contributes information to the estimation of the relative placement of each country-year unit relative to every other one. Again, this placement is what the values of the latent variable mean. Each item contributes to the estimation of these positions. These values are then interpreted as having meaning related to the theoretical concept of interest: respect for human rights.

1.21

The incorrect functional form results in an artificial, strong, positive trend in estimated latent scores even if the actual values of standards-based human rights indicators are replaced by randomly simulated values or are excluded completely

- These simulations create random ordered variables. Because the data are not generated from any underlying model related to the other data, Cingranelli and Filippov have essential just added random noise to the model. There is nothing we can learn from this simulation.

- I discuss this point in the main manuscript and above. It is worth repeating however, that Cingranelli and Filippov have charged Fariss (2014) with failng to assess the assumptions of the two latent variables models developed in his article. To support this claim, Cingranelli and Filippov create random ordered categorical variables and re-estimate the latent variable model with these additional variables in place of the standards-based variables. Cingranelli and Filippov 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. Each of the standards based variables provide meaningful information for the placement of each of the country-year units relative to all the others. Both the models presented in Fariss (2014) make use of variation in all of the standards-based and events-based variables. The changing standard of accountability model accounts for the frequency of many of the categorical values from the standards based variables relative to the declining frequency of the events-based variables. It does this by allowing the item-discrimination parameters to vary from year to year relative to the fixed value of these parameters for the event-based variables. This is a modeling choice based on the theory presented by Fariss (2014) and discussed in the main response to Cingranelli and Filippov.

1.22

However, in Fariss's specification, that is not the case. Much to our surprise, when we replicated his analysis (using Fariss's own computer code) we found that the reported strong upward trend in human *rights was caused entirely by the inclusion of the event-based data concerning mass killings, genocides and political executions by oppressive regimes. No standards-based data were necessary at all. As an illustration, we replicated Fariss's analysis replacing the actual values of four CIRI Human Rights components with randomly generated data and produced an identical trend (Figure 1)*

- This should not be a surprise. In Fariss (2014), there is clear difference in the trend of yearly average for the constant standard latent variable model and the changing standard latent variable model. The difference between the two average trends occurs because of the conceptual distinction in how the two models treat the standards-based variables in relationship to the event-based variables. These differences are clearly reported, and discussed. Moreover, the conceptual reasons for this difference are clearly articulated in the theory. Cingranelli and Filippov prefer a model that shows no change over time but they have not provided an argument about why the human rights documents are produced consistently year after year. Again, the theory in Fariss (2014) is not about the coding process once the human rights documents are produced but rather is focused on the production of these documents. "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" (Fariss, 2014, 308).

1.23

We also repeated Fariss's computations using the 'event-based' indicators alone and, again, we obtained latent scores that also show a strong, improving trend (Figure 2). It is remarkably similar to the trend of the latent scores reported by Fariss (2014: 308; also shown in our Figure 1). In contrast, when we replicated Fariss analysis including only standards-based indicators, there is no visible upward trend in the calculated latent scores (Figure 3). A comparison of Figures 2 and 3 illustrate the overwhelming dominance of the event-based indicators in determining Fariss's human rights scores. This is clear evidence that the standards-based indicators do not have much effect on the latent scores as Fariss *estimates them.*

- As stated in my response to the point above, the differences in the trend of the yearly averages should not be surprising. These differences are discussed at length in Fariss (2014).

- The final sentence in this statement is factually incorrect. All of the indicators provide important information for the estimation of the latent trait. Again, the item discrimination parameters reported in (Fariss, 2014) provide this evidence. The yearly average is based on all the units available in each year. There is substantial variation within years that the authors do not consider.

- Finally, to unpack the difference between the yearly averages for the changing standard of accountability 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. 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 variables enter 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, based on the standard of accountability model, consistently coded from year to year relative to the event-based variables included in that version of the model. I begin with these variables that are constant over time with respect to the relative frequencies of the event-based variables (the ITT torture and ill treatment variable and the CIRI political imprisonment variable) in addition to the new PTS HRW variable, which only spans three years of coverage. I then estimate the constant standard latent variable models, 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 PTS State Department scale, the CIRI torture scale, and the Hathaway torture scale. 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, which is displayed in Figure 9. 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 that begins with the 7 event-based variables displayed in Figure 10, which begins with the 7 event-based variables displaced. Again, all of these models are estimated with fixed item-difficulty cut-points for each versions of the constant standard model.

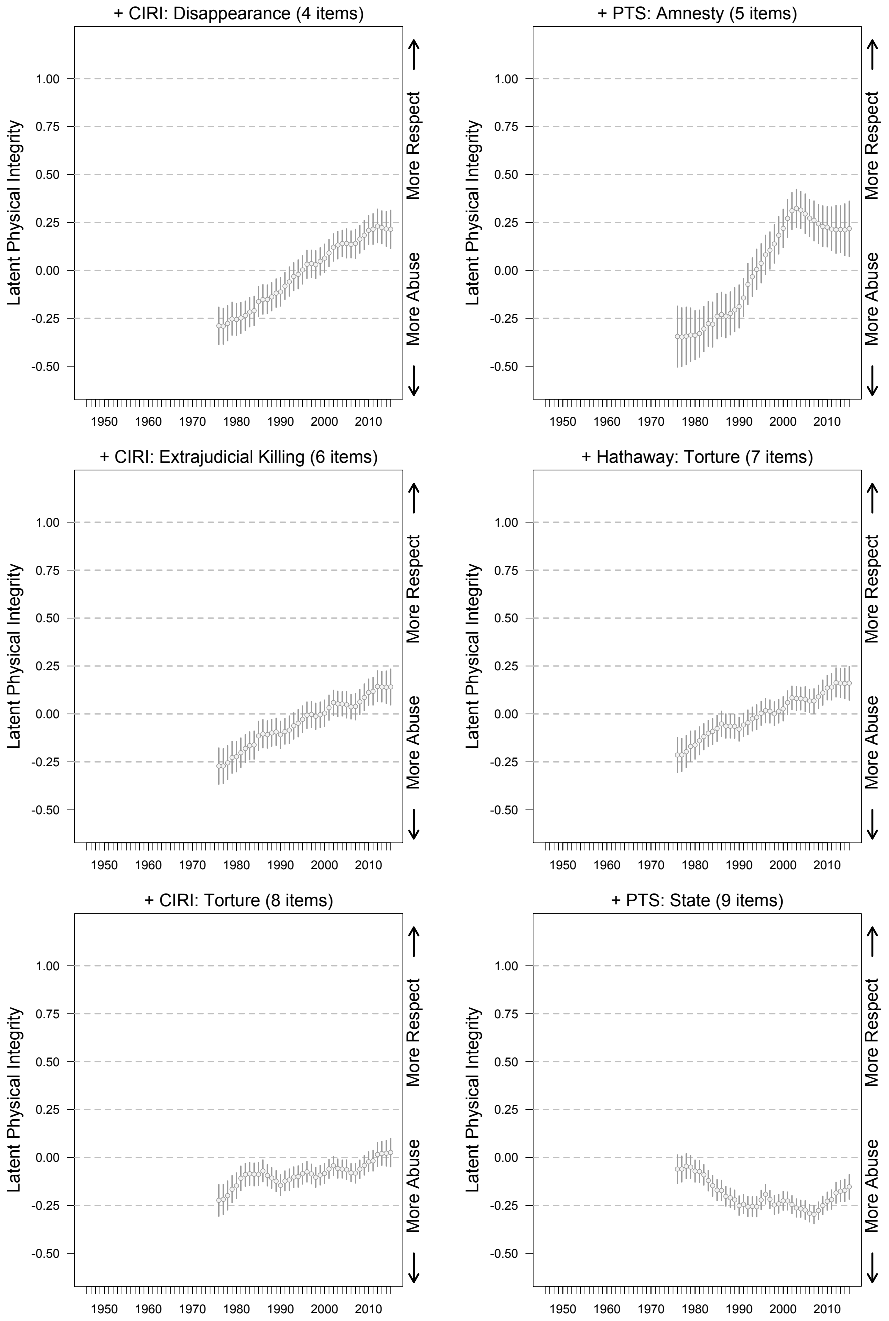

Figure 9: Trends in latent variable estimates for models based on just the standards-based human rights variables over time. All of the models are estimated with fixed item difficulty cut-points (constant standard of accountability). The baselines models begin with the PTS HRW, ITT torture, and CIRI 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.

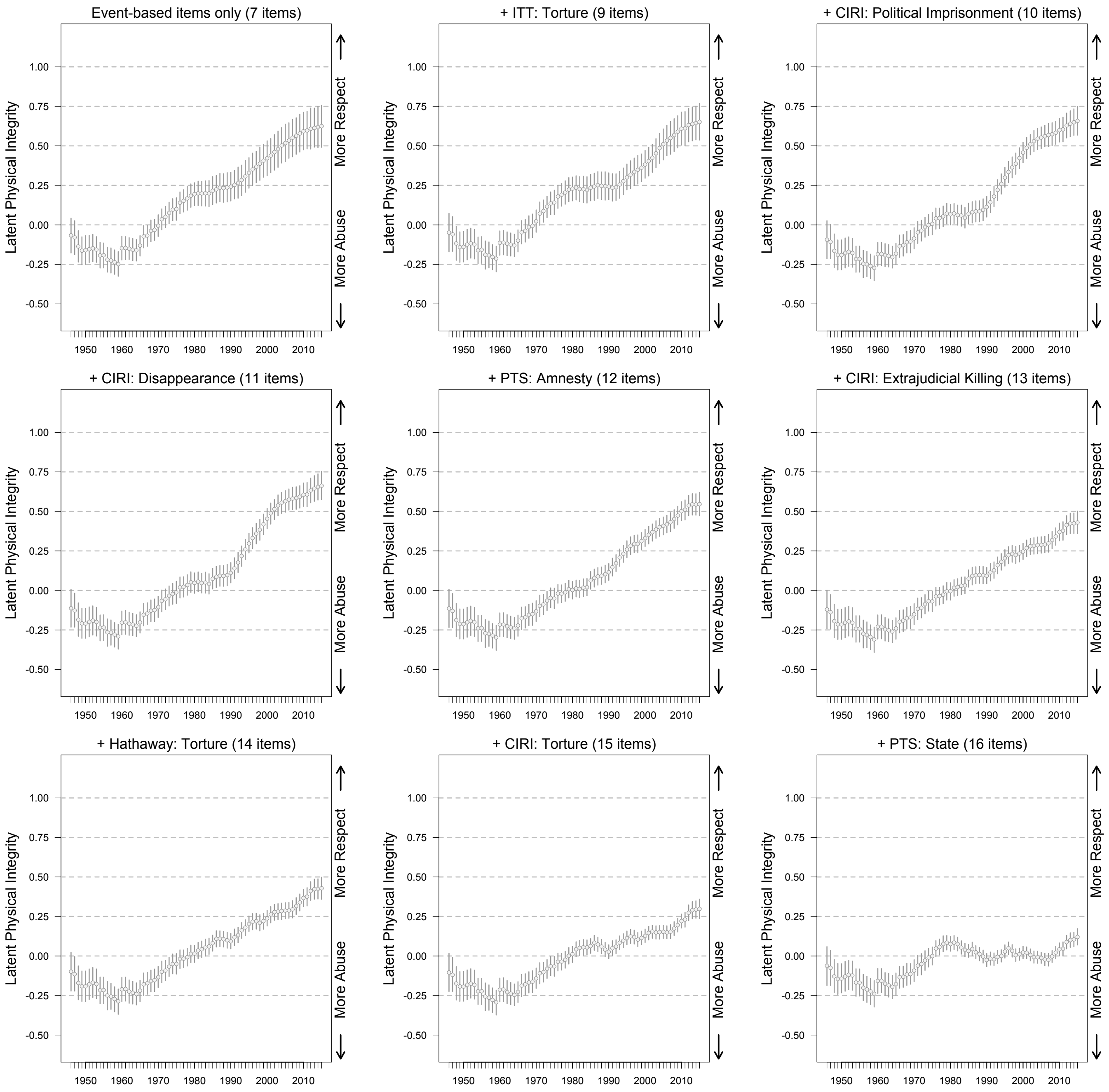

Figure 10: Trends in latent variable estimates for models based on the event-based human rights variables over time with an additional standards-based item. All of the models are estimated with fixed item difficulty cut-points (constant standard of accountability). The baselines models begin with the 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.

1.24

To summarize, Fariss found an improving trend in human rights, because he specified his dynamic IRT model so that instances of mass killings, genocides and political executions determine all the results.

- This statement is factually incorrect and misleading. Each of the variables that enter the model contribute information that is used in finding the best placement for the country-year unit along the latent trait.

1.25

Contrary to accepted procedures for applying dynamic IRT, when Fariss combined the two types of data in a single estimator, he treated the two groups of indicators differently (Fariss 2014: 305-306). He set the "standards-based" indicators to follow an ordered logistic regression with variable intercepts for every year (time-varying item cut points), but the "event-based" indicators were set to follow a logistic regression with a fixed intercept (cut point). This means that the latent variable had to fit actual observations of the "event-based" indicators without allowing a possible adjustment to the intercepts. With the "standards-based" indicators, on the contrary, it was much easier for the algorithm to fit the latent variable as there were several dozen additional parameters (time specific intercepts) that could also adjust. Consequently, variation in the "event-based" indicators drove the estimation of the latent variable. On the other hand, variation in the standards-based indicators was ignored because of the misspecification of Fariss's estimation which allowed for variable intercepts for some indicators but not for others. Thus, it did not matter what the actual values were in the less restricted standards-based part of the Fariss's estimator; only the more restricted event-based part mattered.

- This statement is factually incorrect. Differential item functions (DIF) allows the latent variable model to adjust the values of the latent variable given knowledge of the differences in the process by which the observed information arises. This is what the changing standard of accountability model does.
- In the main response to Cingranelli and Fillippov, I plot the mean estimates for the latent variable in each year, which I reproduce below (see Figure 11). Each of the indicators matter in each

year. This is why the placement of the countries in each year look relatively similar. However, the placement of the center of the distribution in each year is driven entirely by the prior value of 0 for the Cingranelli and Fillippov model. This value is determined by the relationship between the data each year in the other two models. The changing standard of accountability model changes over time because of the declining frequency of the event-based variables relative to the static frequency of several of the standard-based variables from year to year. This is the key difference in the two types of data that lead to this difference over time.

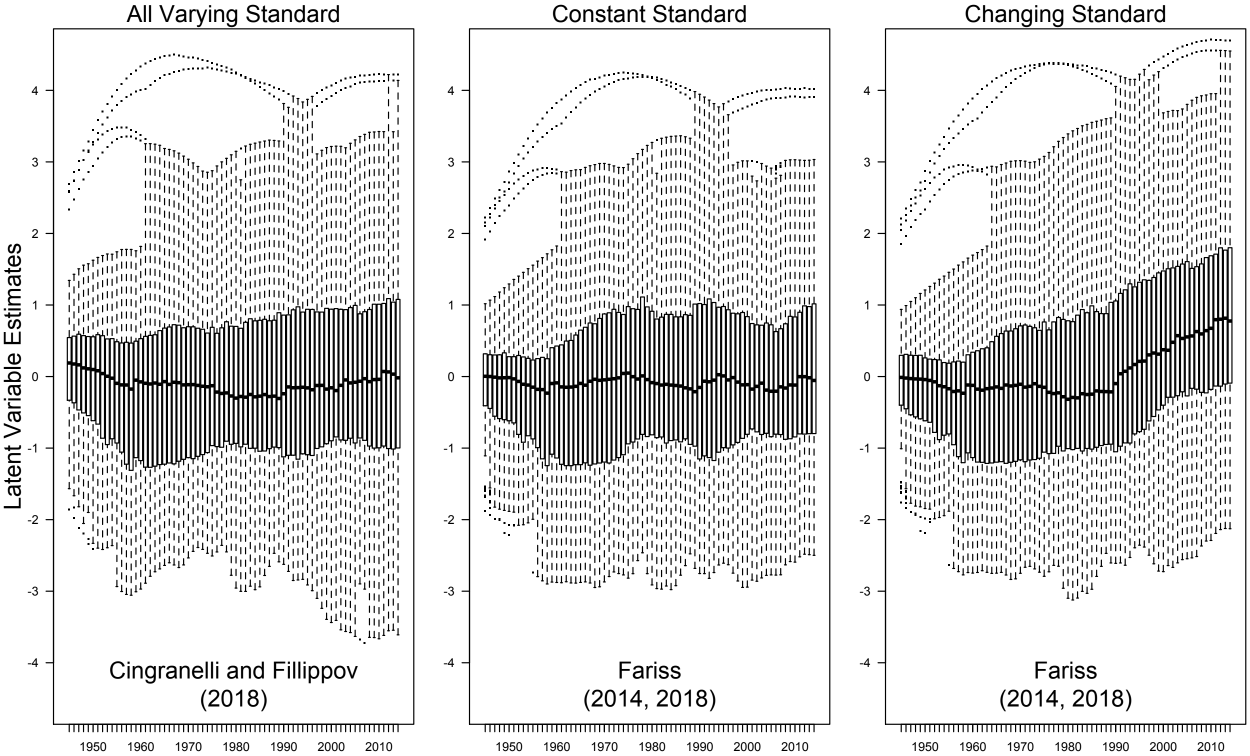

Figure 11: Each panel shows the distribution of point estimates for the human rights latent variable estimates from three competing models. The model proposed by Cingranelli and Filippov (left) produces a very similar trend to the constant standard model from Fariss (2014) (middle) because in the all-varying 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 because of the influence of the standards-based variables. Because model proposed by Cingranelli and Filippov resets the mean estimate for all the countries each year, the mean estimate can never move away from. This is why the median and interquartile range estimates is so consistent from year to year in the left panel. The small changes year to year are due in part to the new countries that enter the dataset in later years.

1.26

The proper way to use dynamic IRT is to treat all observable indicators similarly in their relationship with the latent variable. This is how dynamic latent IRT models were used previously (e.g., Martin and Quinn 2002; Park 2011; Schnakenberg and Fariss 2014; Wang et al 2013).

- A latent variable model must be identified with respect to the parameters the model should be able to estimate and compare.
- If some of the indicators are biased, then the relative placement of one unit relative to others will be biased too. The latent variable values are these placements.
- Again, Differential Item Functions allow the model to adjust the values of the latent variable given knowledge of the differences in the process by which the observed information arises. This is what the changing standard of accountability model does.
- For the constant standard model to be more consistent with reality, the human rights 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.

1.27

A test of the argument about the changing standards of accountability in human rights records should be parameterized in the dynamic model by assuming that all indicators could have variable intercepts. This approach leaves the possibility of a fixed intercept to be endogenously generated in the estimation. When we re-run Fariss's analysis allowing all indicators to have variable intercepts (in all other ways relying on Fariss's original computer code), we obtain the trend displayed in Figure 4. It shows no evidence of human rights improvement. These results are robust to the choice of indicators included in the estimation from four individual CIRI components to all 13 available indicators.

- This model, as describe, is not identified with respect to time. It cannot test for the changing standard of accountability. It cannot account for any changes in the latent variable over time. This is because this model re-estimates the latent variable for the set of units in each year.

- Because the prior distribution for the latent variable is a normal distribution, centered on 0, the average value of the latent trait is forced to be 0 each year for a model in which all the item difficulty parameters (the intercepts or cut-points) vary from year to year. This is because the intercepts or set of cut-points are all estimated for each item for each year of data as if they were independent models.

- The simulation study that I presented in the main manuscript demonstrates that the model proposed by Cingranelli and Filippov is not capable of capturing over time change even if the underlying latent trait does in fact change.

- For any latent variable model to be identified with respect to time, at least one of the item difficulty parameters must be fixed (i.e., estimated as a single parameter for all of the country-year units). Ideally, more than one of these parameters should be constant for all units. Fixing at least one of these parameters to be constant for all the country-year units, allows the mean of the latent trait to vary from year to year, if and only if the frequency of the values for these observed variables change from year to year.

- We can still estimate the model proposed by Cingranelli and Filippov. The model, coincidentally, provides very similar estimates to the constant standard model presented by Fariss (2014) because the mean of the latent variable for this model does not change over time. However, the model proposed by Cingranelli and Filippov does provide some important new concurrent validity evidence that helps us to differentiate its estimates from the constant standard model and the changing standard of accountability model presented in Fariss (2014), which I discuss in Figure 12 below.

- In sum, the model proposed by Cingranelli and Filippov 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 (see the concurrent validity section in the main response for more details). The reason for this is because the model proposed by Cingranelli and Filippov is forced to place the worst cases further down into the negative portion of the latent space as other cases get better conditional on the data since the mean value from year to year is set to 0 by assumption. Because of this modeling choice, the worst cases today are forced further down into the negative portion of the latent variable space but these positions are not relative to other bad cases from prior years.

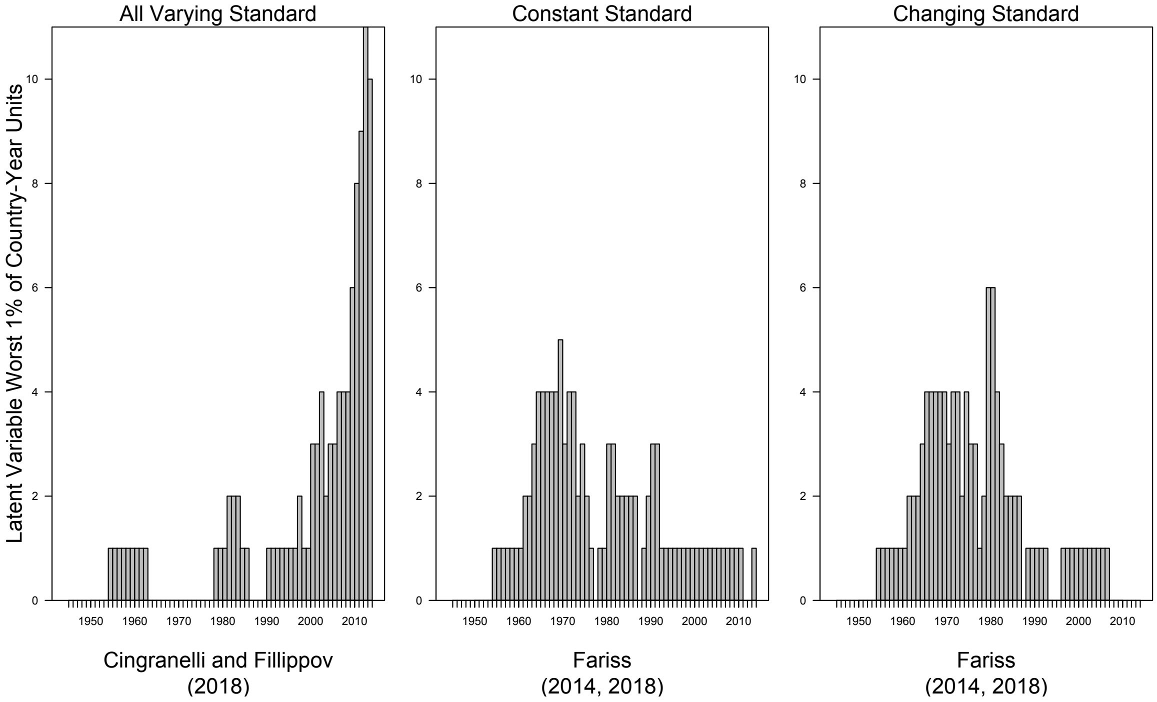

Figure 12: The stacked units in the barplot are the country-years that have the 1% worst scores from three models (all varying model, constant standard model, changing standard model). The all varying cut-point model proposed as the most proper by Cingranelli and Filippov estimates that the worst abusing human rights cases (country-year units) are occurring in the most recent 5-15 years or so. The top 20 worst cases for this model are Sudan 1999-2015; Syria 2013-2015. Whereas the changing standard of accountability model picks: China 1968-1971; Uganda 1976; Afghanistan; 1980-1987; Sudan 1959, 1965-1966; Iran 1982-1985. The constant standard model and changing model are actually in pretty 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 85%. The same indicator for the all varying model correlates with either the constant standard model or changing standard model at about 48% or 39% respectively. The all-varying intercept (cut-point model) 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.

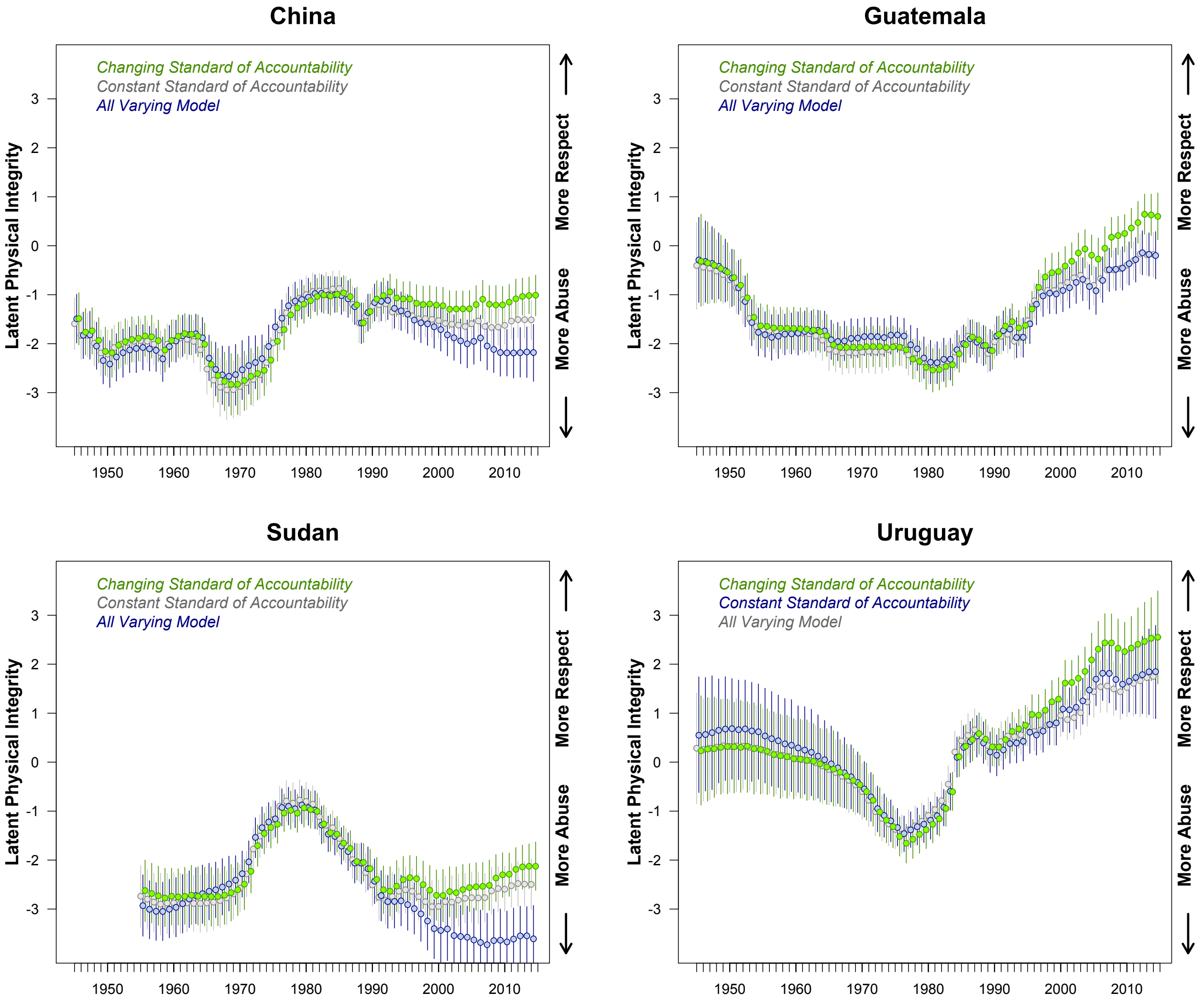

Figure 13: Human rights trends for four countries. In China and Sudan, the worst cases today are forced further down into the negative portion of the latent variable space. These positions are not relative to other bad cases from prior years. The reason for this is because the model proposed by Cingranelli and Filippov is forced to place the worst cases further down into the negative portion of the latent space as other cases get better conditional on the data since the mean value from year to year is set to 0 by assumption.

1.28

Fariss' decision to treat the indicators differently is surprising, because substantively, his breakdown of human rights indicators into two groups is artificial. He makes an assumption, that instances of mass killings, genocides and political executions by oppressive regimes can "act as a consistent baseline by which to compare the levels of the standards-based variables" (2014: 301). In our view, there is no good reason to treat indicators of mass killings, genocides and political executions as more valid, reliable, or in any other way superior to other standards-based indicators.

In comparing mass killings (event-based) data with data derived from coding human rights reports (standards-based), Fariss suggests that records of mass killing events are more trustworthy, because they are "constantly updated," while the standards-based scores are not.

- As cited above, Fariss (2014) states the event-based variables are consistently updated. Most of these variables updated at least three times. To date, the UCDP data has been updated 13 times.
- It is not possible to update the human rights reports, because they are produced by Amnesty International and the US State Department each year. These monitoring organizations do not go back and updated old reports when new information becomes available or when previously unreported types of abuse become a topical focus for observers.

1.29

More specifically, he argued that, in the case of standards-based data, "the older reports are not updated or revised even if new information about specific repressive events is obtained over time..." (p. 303). Actually, the US Department of State and Amnesty International reports are "revised" and updated since they sometimes add information clarifying a situation that was unclear in the previous year or few years. Updating of this type is not common—perhaps one or two countries each year, on average. However, we see little evidence of practical revising of mass killing events data either, especially because the data are only recorded as either ZERO or ONE. Moreover, three of the five events projects Fariss used in his study have not been updated since 1990, so the data included in those projects can never be revised.

- Fariss (2014) actually suggests that the event-based variables "are a valid representation of the historical record to date" (298).

- Each of the event-based variables is accompanied by coding documentation which specifies the period in time when updating and revision occurred. For example, Rummel discusses the process by which he periodically updated the event-based information used in his articles and books.[3] Again, the UCDP data has been updated 13 times to date. See each of these code books and the the other response (Fariss, 2018*a*) to Cingranelli and Fillippov for additional details about these event-based variables.

1.30

Even more importantly, both sets of indicators are standards-based. In practice, once scholars move to code the events of atrocities, they necessarily make judgment calls that rely on some "standards' which we might refer to as "coding rules".

- The argument in Fariss (2014) about the changing standard of accountability is *not* about the coding rules, the operational protocol, developed by the CIRI or PTS academic teams. The standard of accountability is the core concept from a theory about how the organizational structure of human rights monitoring organizations produce information about state behaviors over time (Fariss, 2014). Despite the claims to the contrary by Cingranelli and Filippov, it is not a theory about the operational definitions and coding procedures that take annual human rights rights reports and categorize them into data by academics. It is a theory about organizations and how these organizations change the procedures they use to produce human rights reports each year.

- When creating a measurement procedure, every scientist endeavors to consistently apply it so that the data obtained from the procedure is comparable over time or across contexts. This is necessary requirement for science. It is not a necessary requirement for the activities of monitoring organizations. Moreover, these monitoring organizations are quite explicit that they are not making

[3] See the discussion in the preface of Rummel's book *Death by Government: Genocide and Mass Murder in the Twentieth Century* (1994, 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`.

documentation that they expect to be comparable year to year or even across contexts within year (Eck and Fariss, 2018). The theory of the changing standard of accountability is about how these organizations produce information not how scholars code information from pre-existing source material.

1.31

The only possible difference between the two types of indicators that we can discern is that the CIRI and PTS data generation projects use somewhat better defined and more stable standards than the criteria used to create dichotomous indicators of mass killings, genocide, and democide.

- It is unclear what criteria Cingranelli and Filippov believe makes the operational protocol from the CIRI and PTS projects better defined than those from the other repression measurement projects.

1.32

For example, Harff and Gurr (1988), who developed one of the most respected data sets on genocides and other types of mass killings, acknowledge that their coding was standards-based. Like the PTS and CIRI coders, they also relied on the annual Amnesty International and the US Department of State Reports, among other sources, to identify their cases. They say that they developed "detailed operational guidelines" distinctive from other scholars of mass killings and specifically" the list of episodes includes only those in which (a) many noncombatants were deliberately killed, (b) the death toll was high (in the thousands or more), and (c) the campaign was a protracted one" (Harff and Gurr, 1988: 365).

- The theory of the changing standard of accountability is not about the coding procedures developed and implemented by the different human rights coding projects.

- Again, the standardization of the documentation process used to produce the human right reports is not under the control of the PTS or CIRI project. It is under the direct control of the organizations themselves. This is the standard of accountability. It is a normative standard that 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).

1.33

However, in her 2003 article published in this Review, Harff (2003: 60) admitted: "some of the figures are little more than guesses."

- This quote is taken out of context.

- Specifically, when Harff (2003) states that "some of the figures are little more than guesses", she is referring to the estimated number of victims, which is a column in Table 1 of her article. She is not referring to the occurrence of the genocide or politicized, which is a binary, categorical variable. Harff and Gurr (1988) make a similar statement about the same range of estimates on page 365 of their article. Again though, this statement is about the estimated count. The latent variable models presented by Fariss (2014) use the binary categorical variables from Harff (2003) and Harff and Gurr (1988). Fariss (2014) is quite clear about this distinction.

- The authors also select a quote from Poe and Tate (1994) about count data and again misuse it. The particular passage by Poe and Tate (1994) is not about Rummel (1994), which Cingranelli and Filippov imply. In the endnote section of their paper, Poe and Tate (1994) are discussing the critique that using counts from newspaper sources might led to misleading comparisons. Specifically, Poe and Tate are discussing use of data from Taylor and Jodice (1983), which is also data used in Fariss (2014). But, as with the Harff (2003) data, these event counts are also turned into a binary categorical variable for the latent variable models in Fariss (2014). Does evidence exist that supports whether or not such an event took place in a specific country-year? This is what the categorical variable tells us. None of the latent variable models in (Fariss, 2014) make comparisons of the event count data available from some of the event-based datasets.

1.34

Rummel (1994) used the least restrictive coding rules. He recorded more mass killing events than any of the other scholars whose data Fariss used, and his list of mass killings has the strongest effect on Fariss's scores. Rummel even recorded the United States as committing mass killings based on civilian wartime deaths in Korea and Vietnam. Poe and Tate noted that, based on such records of mass killings, the United States was a more repressive society ... than the Soviet Union under Stalin? (1994: 868).

- This quote from Poe and Tate (1994) is taken out of context (see details above).
- It is not clear from the text how the authors calculated this relative effect size because it is not consistent with the size of the item discrimination parameters presented by Fariss (2014).
- This statement is a critique of Rummel. Cingranlli and Filippov make the same claim in their earlier critique of Fariss (2018*b*), which is responded to in Fariss (2018*a*). Specifically, Fariss (2018*b*) discusses this deviant case of the US in 1953 at length and also defends of the Taylor and Jodice (1983) and Rummel (1994) data projects.
- It is important to note however, that Rummel did count both international and domestic deaths attributable to the US government. The binary coding for the US is a 1 based on Rummel is for the domestic deaths. This choice would only be problematic if the case identified by Rummel only included government killings outside of the state in question. Again see Fariss (2018*a*) for more details on this particular case and more details on the coding of these events based variables.

1.35

To summarize, then, the indicators of mass killings, genocides and politicides are no more valid than other data widely used in the human rights subfield. If anything, some of the events datasets Fariss used are less valid and reliable. Glossing over obvious weaknesses in those indicators, Fariss used the dynamic IRT model inappropriately to privilege them over more conventional, widely used indicators. As a result, many of Fariss's scores fail face validity. For example, one of the strongest findings in the human rights literature is that, other things being equal, democratic countries and economically developed *countries have better records of human rights performance than others. Despite these well-established facts, many of Fariss's scores place authoritarian, less developed countries higher than well-established democracies.*

- Fariss (2018*a*) addresses each of these points in detail in the first exchange with Cingranelli and Filippov.

- The period of time for which the CIRI standards-based variables and the Rummel event-based variable are available is 1981-1987. During this period, the event-based binary variable based on the Rummel event count data has spearman correlation coefficients that are similar in size to the CIRI variables as the other CIRI variable have with each other. The Rummel binary event-based variable has the highest correlation with the standards-based CIRI variables compared to the other event-based binary variables. These categorical variables are tapping into the same underlying concept: respect for physical integrity rights.

	Rummel	TORT	POLPRIS	KILL	DISAP
Rummel	1.00	-0.39	-0.44	-0.53	-0.46
TORT	-0.39	1.00	0.59	0.55	0.43
POLPRIS	-0.44	0.59	1.00	0.52	0.41
KILL	-0.53	0.55	0.52	1.00	0.57
DISAP	-0.46	0.43	0.41	0.57	1.00

- Below is a graph that displays the average level of the latent human rights variable over time for democratic and non-democratic country years from both the changing standard and constant standard models. If the latent variable model was systematically placing authoritarian states above democratic ones, an issues of concurrent validity, then these averages would overlap.

- Again, deviant cases such as Sweden with respect to its scores on the CIRI torture scale are useful entry points for discovering new concepts that might influence how information is obtained by monitoring organizations (Eck and Fariss, 2018; Lijphart, 1971; Seawright and Gerring, 2008).

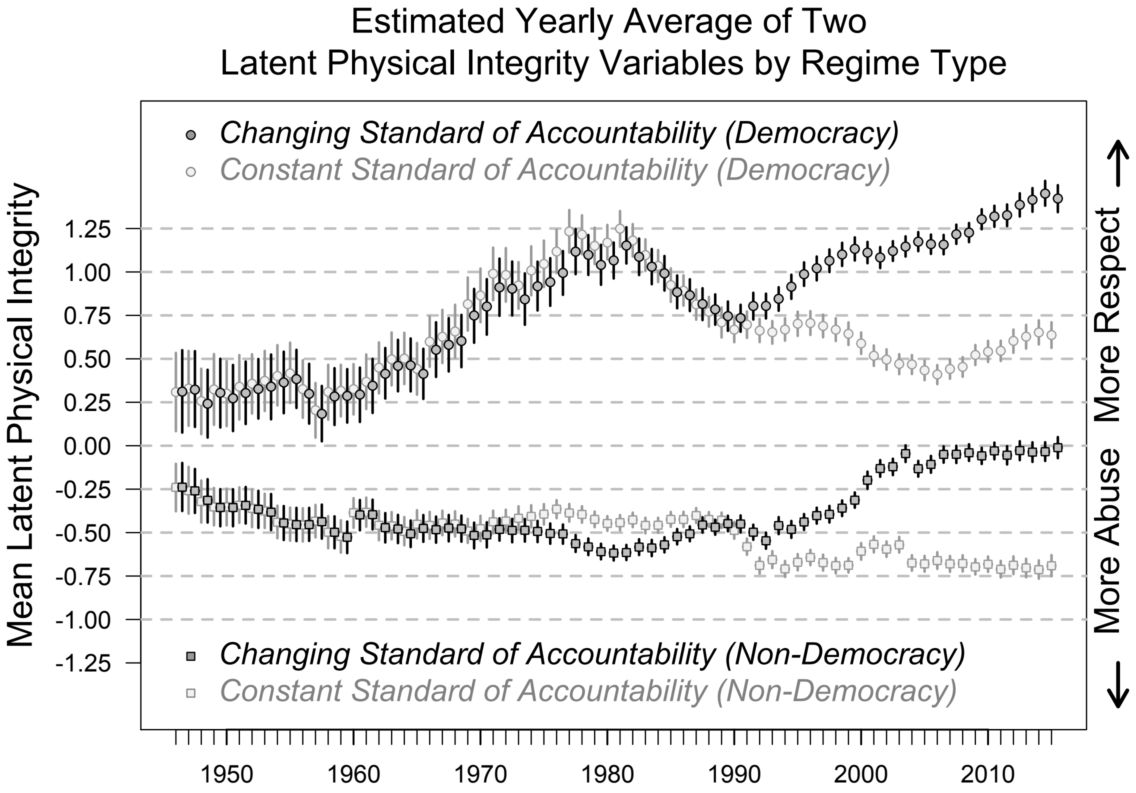

Figure 14: Modified from Fariss (2018*a*): The graph displays yearly mean and credible intervals for these same variables across democratic and non-democratic states as measured by Polity IV (values of 6 or greater). Only the latent variable estimates that assume a changing standard of accountability show improvement for either type of country-year. Without the assumption of the changing standard of accountability, one must believe that the level of human rights in just the set of democratic states has been decreasing since a high point in the early 1980s. It is more likely that the standard of accountability is improving as monitoring agencies look harder for abuse, look in more places for abuse, and classify more acts as abuse. See Fariss (2014) for additional details.

1.36

Fariss's theoretical argument and statistical model do not give a reason to expect the cross-sectional rankings of countries as indicated by their relative dynamic latent scores to be significantly affected, but they are. The model assumes that the standards of accountability in the human rights reports by the US Department of State and Amnesty International change for all countries in the same way over time. The rate of change may be greater during some time periods, but Fariss does not maintain that the changing standards of accountability should have different effects depending upon the particular country. Despite having no foundation in his theory, we do find substantial cross-sectional and over time disturbances in his latent scores. These disturbances are severe because the binary data on mass killing events strongly influence the computation of the latent scores. The consequences of such disturbances are the most significant for thousands of cases between 1946 and 1975. Fariss extrapolated those early scores from very limited records. For the vast majority of country-years between 1946 and 1976, there were no recorded genocides (95% of all country-years), political executions of dissidents (90%) or mass killings (77%, according to Rummel), so the actual codes before extrapolation were ZERO. Yet all country-years were assigned different latent human rights scores including the very large group of country-years for which there were only actual scores of ZERO.

- I addressed this point on extrapolation in detail above. In review: when fewer items are available such as the 1946-1975, there is greater uncertainty about where to place the unit. This uncertainty is captured by the standard error of the latent estimate. I provide a graph that visualized this here in Figure 20.

- The country-year units that receive only good scores (0s) for the event-based variables have the most uncertainty associated with their unit specific estimates. This means that the relative placement of these units relative to the others in the model are the most uncertain. These units could have mediocre or good records but the latent variable model does not have sufficient information to differentiate these two types. It is therefore incumbent on the researcher to include the uncertainty information in any model that includes the latent human rights variable, which is a point that Schnakenberg and Fariss (2014) discuss at length. Fariss (2014) discusses this at length in section M of the supplementary appendix that accompanies this article.

- See also Figure 11 above, which shows the cross sectional variation of the mean estimates for the three latent variable models presented in the main response to Cingranelli and Fillippov.

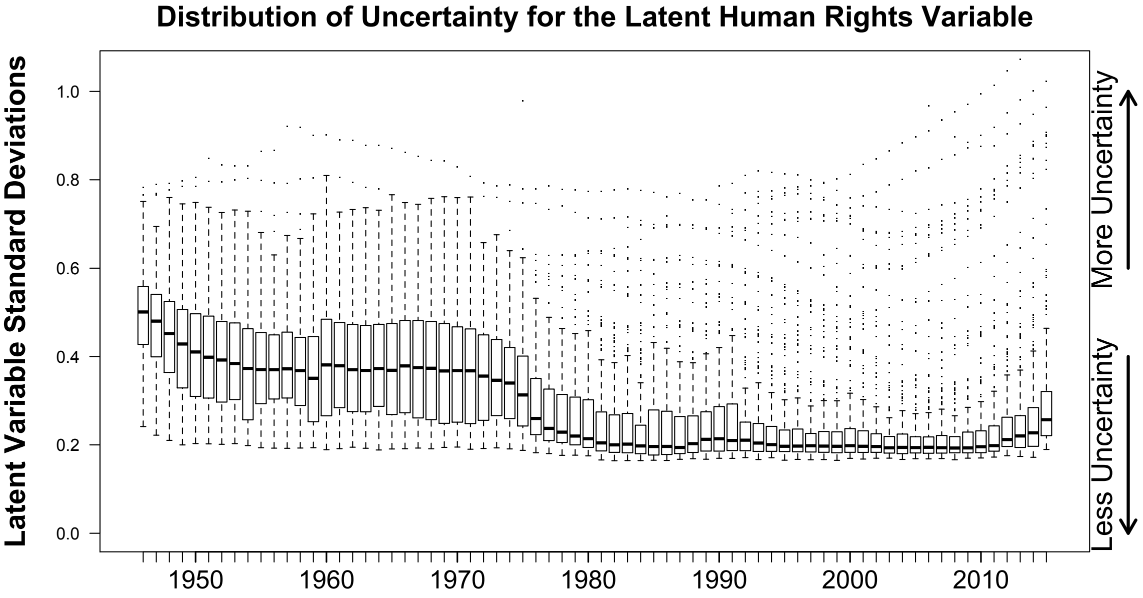

Figure 15: Modified from Fariss (2018*a*): The yearly distribution of the standard deviations from the latent variable estimates from 1946-2015. Though not every one of the repression variables is measured for each country-year unit, the latent variable model is able estimate a value of the latent variable for each country-year unit using the observed variables that are available. As this graph illustrates, the level of uncertainty for each country-year unit is in part a function of the availability of the observed variables. Thus, there is more uncertainty in earlier years and importantly this uncertainty information can be incorporated into standard statistical analyses (Schnakenberg and Fariss, 2014). As new repression variables are incorporated into future versions of the latent human rights model, these estimates will decrease, conditional on the relative quality of those new variables. See Fariss (2014) for additional details.

1.37

Fariss's new method for calculating dynamic latent scores has not been properly evaluated. Its pitfalls and biases are unknown. We have identified serious problems already. The strengths and weaknesses of the previously used versions of dynamic IRT have been widely discussed in the literature. We can find no other example of using dynamic IRT where the cut points for some variables are allowed to vary but for others they are not. Without these unique specification choices, Fariss's (2014) results do not emerge

- There are actually a number of examples: see Caughey and Warshaw (2016) or Hare et al. (2015) for recent examples. For a thorough and helpful review on the state of political science methods work using IRT model see (Imai, Lo and Olmsted, 2017). See Gelman and Hill (2007) for a general discussion of the Bayesian approach to estimating hierarchical parameters.

- The average improvement over time, which is the only empirical result in question in this critique, is replicated by an updated analysis presented in the main manuscript and three other latent human rights variables from the Variety of Democracies project. Cingranelli and Filippov have not yet addressed these other results from Fariss (2018*a*). See Figure 8 above for these similar results.

1.38

It is a standard in the literature that new statistical procedures should be evaluated using simulation methods (e.g., Monte Carlo simulation), specifying the data generating process and verifying that the model returns sensible results. Fariss (2014) does not present the results of such simulations. In the future before any new latent dynamic scores are introduced, the techniques of their calculations should be examined and tested properly.

- Indeed, such a simulation study demonstrates that the latent variable model specification proposed by Cingranelli and Fillippov is not identified with respect to time. The model proposed by Cingranelli and Filippov is similar to modeling each year of data as a separate model and then combining the estimates back together. Instead of letting the latent variable model use the distribution to arrange all of the country-year units relative to one another for all years, the all-varying cut-point model is reseting the distribution each year and only arranging the country-year units relative to one another within one year. This is why the estimates are not useful for making comparisons from one year to any other year though these estimate are still are suitable for making comparisons within a given year.

- The parameterization of the function for each of the observed human rights variables in the latent variable models developed in Fariss (2014) are similar to other models of ordered and binary data. Researchers in American political are making similar modeling choices (e.g., Caughey and Warshaw, 2016; Hare et al., 2015). The specification of the differential item functions in Fariss (2014) are supported by a rigours theory about the changing standard of accountability which I have discussed extensively in the responses above and in the main manuscript.

1.39

Unfortunately the human rights scores Fariss produced already are widely used. Some scholars now treat his measure as the new standard. The use of his erroneous data to re-examine well established findings in human rights studies is a serious mistake. It will lead to a substantial detour in theory building and in evidence-based policy making. This is the main reason why we present this critique.

- For the past fifteen years or so, scholars have puzzled over the why UN human rights treaties were negatively correlated with measure of human rights, primary the CIRI, PTS, and Hathaway human rights data (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. The negative patterns are not valid because the data did not account for changes in the source material used to generate the categorical data. 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 (see Figure 8) and by other scholars working on understanding the relationship between law and human rights (e.g., Dancy, 2016, 2017; Sikkink, 2011, 2017).

- 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). Now there is reason for both hope, new theorizing, and new data collection. Indeed, a new and growing literature is now working to understand and document how these mechanisms work to causes the standard of accountability to change 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; Clark and Sikkink, 2016; Fariss et al., 2015; Park, Greene and Colaresi, 2017). This is the promise of a science of human rights (Schnakenberg and Fariss, 2014).

1.40

We recommend that neither scholars nor policy analysts should use Fariss's scores for studies of physical integrity human rights. We have shown that Fariss did not use the existing physical integrity data in a meaningful way when he computed his scores. Computationally, the pattern of Fariss's dynamic latent scores mainly reflects the changes in frequencies of extreme events of mass killings and genocides. Privileging mass killing and genocide indicators may be a researcher's preference, but it makes Fariss's scores a poor source of inference about the dynamics of the excluded categories of variables (i.e., torture, extrajudicial killing, political imprisonment, and disappearances). At best, he has created a dynamic latent score that reflects the decreasing incidence of mass killing events over time.

- The reanalysis of the latent variable model presented by Cingranelli and Filippov does not support these claims. Cingranelli and Filippov misunderstand the theory of the changing standard of accountability and privilege estimates from a latent variable model that is not identified with respect to time. This is because the model proposed by Cingranelli and Filippov is similar to modeling each year of data as a separate model and then combining the estimates back together. Instead of letting the latent variable model use the distribution to arrange all of the country-year units relative to one another for all years, the all-varying cut-point model is reseting the distribution each year and only arranging the country-year units relative to one another within one year. This is why the estimates are not useful for making comparisons from one year to any other year thought still are suitable for making comparisons within a given year.

- Cingranelli and Filippov have provided little evidence to support their argument that the standard of accountability is not changing and that human rights are not improving over time.

1.41

For those who choose to use Fariss's dynamic scores despite our critique of them, we warn against using latent scores as dependent variables in conventional regression analysis. This can result in inconsistent or severely biased estimates. Instead, when using latent scores, it is necessary to use more advanced statistical techniques such as simultaneous equation analysis, data simulation or multiple imputations (Bolck, Croon and Hagenaars 2004).

- It is not quite clear what the methodological suggestion here is. One of the suggestions in the article by Bolck, Croon and Hagenaars (2004), is that researchers should not treat estimates of latent variables as population estimates or perfectly observed. What this means is that latent variables are estimated with uncertainty and that this uncertainty should be incorporated into subsequent analyses that use the latent variables as independent variables. I discuss this in several of my responses above. It is also discussed at length by Schnakenberg and Fariss (2014), particularly in relationship to the CIRI additive index which itself is a latent variable. The CIRI additive index, by assumption, assumes perfect precision and equal weighting of each of the observed items. Building on research by Mislevy (1991), Schnakenberg and Fariss (2014) provide suggestions similar to the recommendations of Bolck, Croon and Hagenaars (2004). Specifically, Schnakenberg and Fariss (2014) suggest incorporating the uncertainty from latent variable estimate using the multiple imputation equation formula from Rubin (1987). Fariss (2014) also discusses this in Appendix section M.

- Fariss (2014) also provides several additional methodological suggestions for using the latent variable estimates or the original categorical variables from PTS or CIRI:

  "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 (Appendix 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" (314).

1.42

Finally, we agree with Fariss that regardless of possible changes in standards of human rights recording, widely used indicators of human rights abuses (e.g., CIRI and PTS) are useful for comparing state behaviors in the same year (Fariss 2014: 303). Thus, a practical way to address concerns about the potential changing standards of accountability in human rights data is to check the robustness of cross-national time-series results with cross-sectional estimations.

- On page 303, Fariss (2014) states that "[t]he 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." This is a statement about the comparability of the information in the reports, not about the statistical comparison of the categorical values from the CIRI or PTS projects.

- If the standard of accountability is changing, then this process needs to be accounted for in models of human rights over time. If a theory specifies over time change, then cross sectional comparisons are not useful robustness checks.

- As I mentioned in the comment above, Fariss (2014) provides several suggestions for using the latent variable estimates or the original categorical variables from PTS or CIRI. See section K in the supplementary appendix that accompanies the article by Fariss (2014).

- Overall, Cingranelli and Filippov have provided little evidence to support their argument that the standard of accountability is not changing and that human rights are not improving over time.

2 Line by Line Response to Cingranelli and Filippov version 2 (new version)

2.1

Scholars and policymakers in the human rights subfield are now facing a significant disagreement. Those emphasizing distinctive types of evidence are reaching different conclusions about trends in human rights and about a variety of other questions relevant to scholarship and policy making. Fariss (2014) provoked the divide when he suggested a novel statistical approach to reevaluate human rights indicators. His new scores showing improving global trends in human rights are at odds with trends in previously used measures (Cingranelli and Richards 2010; Wood and Gibney 2010).

- Important disagreements over the efficacy of human rights law and human rights change have been present in the academic literature for some time. Prior to Fariss (2014), for example, see the discussions in Brysk (1994), Clark and Sikkink (2013) and Goodman and Jinks (2003). Fariss and Dancy (2017) review the evolution of these debates over the last several decades.

2.2

Fariss's new scores encourage scholars to re-examine many research findings accumulated by the subfield. Since his scores, on average, increase over time, they are likely to be correlated with many variable that also increase over time such as treaty ratifications, degree of globalization, the degree of economic inequality within nations, and democratization. Fariss (2014) has already challenged previous results on the effects of human rights treaty ratification. Future studies using his scores are likely to produce many findings that conflict with previous results.

- This is an important part of the scientific process. If an existing measurement procedure produces biased results, then such a measurement validity issue needs to be addressed.

2.3

Fariss's model specification and results were strongly affected by his assumptions.

- See my responses in sections 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6.

2.4

Mass killing events (such as genocide) and lesser human rights violations are indicators of the same underlying variable respect for physical integrity human rights.

- These assumption of unidimensionality is consistent with a large body of literature and empirical evidence. For more details see response 2.29 below.

2.5

Incidents of mass killing are recorded more accurately than lesser violations.

- Fariss does not argue that mass killing are recorded more accurately than other forms of violations. The argument in Fariss is about the documentary evidence itself and distinguishes between the direct categorization of the documentary evidence (standards-based variables) and the indirect 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 than other violations when they are occurring. 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 (e.g., Eck and Hultman, 2007; Harff, 2003; Harff and Gurr, 1988; Rummel, 1994; Taylor and Hudson, 1972; Taylor and Jodice, 1983).
- See section 1.28 for more details about the updating practices of the event-based variables.

2.6

Since there has been a substantial decline in the records of mass killings (Figure 1), other indicators of human rights also should reflect an improving trend. If they do not, it is because of the "changing standards of accountability" (S1 varies) in human rights reports of lesser human rights violations.

- Fariss (2014) argues that the distinction between the two types of variables — event-based and standard-based — is based on the processes by which the information about human rights abuses is produced. The theory is particularly focused on the information production process that takes place through the publication of the yearly human rights reports (e.g., Fariss and Tyson, 2018). This process occurs before the information is used by political scientists to categorize human rights. The academic coding process is not the theoretical focus in Fariss (2014). For the categorical comparisons of the PTS and CIRI data to be valid, these academic teams rely on the consistent application of the same standard year after year by the human rights monitoring organizations when they produce their reports. If the standards these organizations use to document human rights change over time, as I argue they do, then the temporal pattern in the CIRI and PTS data will be biased over time. I return to this point several times below and in the main response. See section 1.6 in particular.

2.7

There has been no change in the 'standards of accountability' in records of mass killings.

- This is incorrect. The standard of accountability is about the process by which documentary evidence is produced. It 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 a normative standard that 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. Unlike the standards-based variables (e.g, CIRI, PTS, Hathaway, and ITT data project), 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 coding and 0 otherwise. For the standards-based variables, documents are directly coded. For the event-based variables, the documents are indirectly coded because the documentary evidence is taken from multiple sources and used to look for evidence that a particular type of repressive event occurred. These are fundamentally distinct processes.

2.8

Indicators of lesser human rights violations that do not reflect an improving trend should be corrected to remove distortion due to the difference in the changes in the standards of accountability (S1 - S2) between the two types of records.

- I assume that Cingranelli and Fillipov are using the term “distortion” as a synonym for bias in the latent estimates as opposed to variance or uncertainty in the latent estimates. I consider the trade off between bias and variance for several related latent variable models in an extensive simulation study presented in Reuning, Kenwick and Fariss (2018).

- This is correct, if the underlying source material is changing, then the categorical indicators derived from that source material is potentially biased and should be corrected. The declining frequency of the event-based variables relative to the static frequency of several of the standard-based variables from year to year is the key difference in the data and the one that is supported by the theory of the changing standard of accountability, which motivates the new latent variable model for human rights that incorporates this theoretical concept. See section 1.6 in particular.

2.9

The crucial assumption is that there has been no change in the standards of accountability in records of mass killings. An alternative, less restrictive assumption is that there also has been a change in the standards for recording mass killings (S2 varies). However, a model based on this assumption would produce an estimation of no improvement in human rights latent scores, because it would assign lesser weights to indicators of mass killing.

- The model that Cingranelli and Fillpov prefer is not capable of making comparisons over time. See the main response to Cingranelli and Filippov and section 1.27.

2.10

Fariss's assumptions, and, crucially, that S2 does not vary, led him to use a statistical technique that heavily weights rare "event-based" incidents of mass killing such as genocide, discounting "standards-based indicators" of "lesser" and more common human rights violations such as torture and political imprisonment. Unfortunately, the model weighted mass killings so heavily that the increase in the proportion of countries with no mass killings (beginning in the mid-1970s) closely mimics the pattern in Fariss's latent scores (Figure 1).

- This statement is factually incorrect. I have addressed related claims about the vales of the parameters throughout the main response and in this appendix. Specifically, see sections 1.2, 1.3, 1.4, 1.5, and 1.6.

- None of the latent variable models considered by Fariss (2014) or in this response discount any of the available information captured in the categorical human rights variables. The relative information content of each of the observed human rights variables is estimated in relationship to each of the other observed items (Schnakenberg and Fariss (2014) discuss this point at length). Most of the within-country variance for each of the latent variables is determined by the standards-based variables (see sections in 2.18 and 2.19 for additional details on this particular point).

- The particular claim that the latent variable model over-weights the event-based variables is in need of additional clarification. When Cingranelli and Filippov are referencing the weighting of a particular item, they are discussing the item-difficulty parameters (i.e., the intercepts), not the item-discrimination parameters (i.e., the slopes or item-weight). I discuss what each of these parameters does in the context of the particular probability model that relates the latent variable estimates to the observed human rights data.

- The item-difficulty parameters (i.e., the intercepts) determine the position of the inflection point of the cumulative density function of the logistic function for each observed item.

- The item-discrimination parameters (i.e., the slopes or item-weights) determine the slope and therefore the spread or shape of the cumulative density function of the logitistic function over the range of the latent variable estimate.

- These two item parameters, in concert, and for each observed item, transform the estimated latent variable into probabilities that are associated with the categorical values of the observed items.

- The intercept or the cut points of the latent variable model account for the relative frequency of each of the observed items, conditional on the value of the latent variable itself. The item-weights, which are slope parameters from the logit or ordered logit link functions transform the latent variables estimates into probabilities that are associated with each of the observed human rights variables. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along of the latent variable that each of the country-year units will probabilistically occupy. Larger item weights — larger logit or ordered logit coefficients — represent increasing precision in the placement of the units along the latent trait when they take on a specific value of the observed item. As the coefficient approaches infinity the logic curve begins to approximate a step function. The inflection point, the point at which the step occurs, is the position along the latent variable that country-year units will be placed above or below conditional on the observed value of the particular observed item under consideration. Larger item-weights are associated with the information content of the particular value of the observed variable in relative comparison to the other observed traits as conditioned by the estimate of the latent trait.

- Figure 16 and figure 17 display this visually below for the event-based variables. The event-based variables are providing probabilistically similar levels of information across the two models when determining the placement of the country-year units along the latent dimension. Below in section 2.15, Figure 18 displays this information graphically for one of the political terror scale variables.

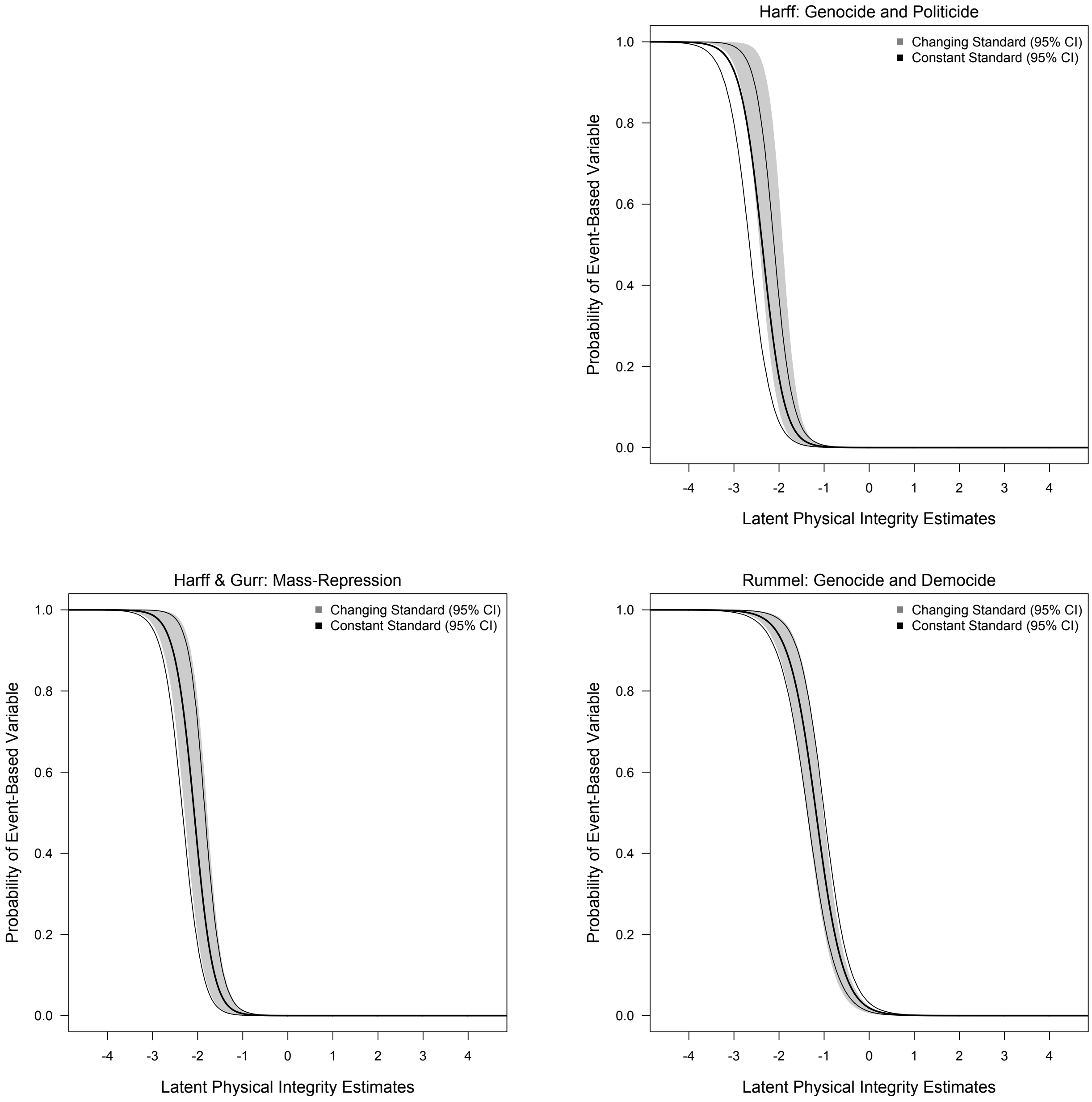

Figure 16: Item-response curves for the event-based variables. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along the latent variable that each of the country-year units will probabilistically occupy. Larger item weights — larger logit coefficients — represent increasing precision in the placement of the units along the latent trait when they take on a specific value of the observed item. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (black) and constant standard models (grey).

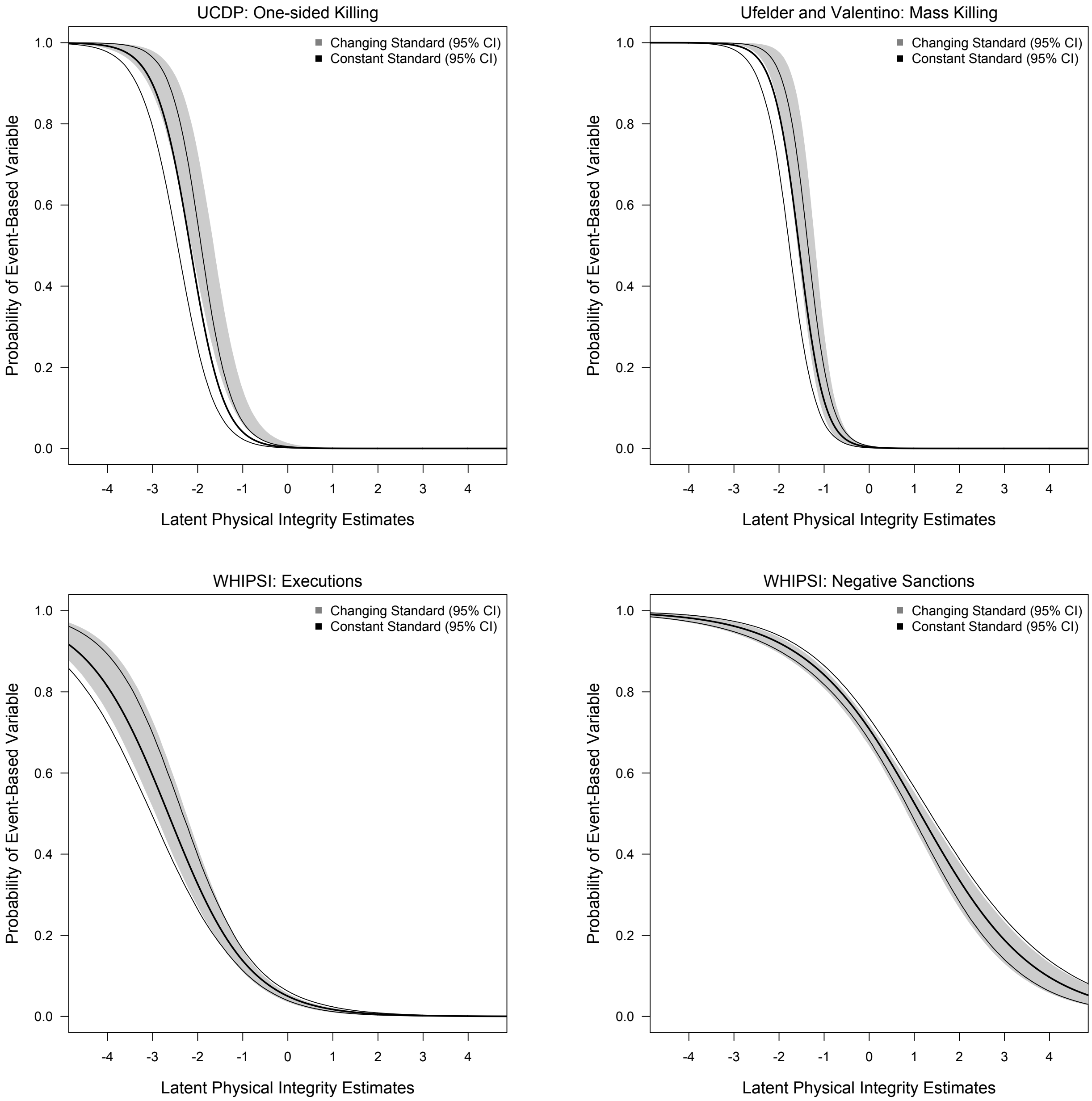

Figure 17: Item-response curves for the event-based variables. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along the latent variable that each of the country-year units will probabilistically occupy. Larger item weights — larger logit coefficients — represent increasing precision in the placement of the units along the latent trait when they take on a specific value of the observed item. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (black) and constant standard models (grey).

2.11

Moreover, as shown in Figure 2, trends in latent scores produced only from records of mass killing trends are hardly distinguishable from Fariss's latent scores. A similar trend also can be generated from mass killing events combined with random numbers substituted for the actual values of lesser violations (Figure 3). Thus, the model he chose would have produced the appearance of an improving trend in human rights between 1950 and 2010 no matter what the records of lesser violations had been. Fariss did not discover an improving trend in human rights. His modeling choice predetermined the trend.

- This statement is factually incorrect.

- See section 1.23 for a discussion of several alternative latent variable models that show the same positive trend in human rights improvement over time for models based on just the standards based variables. For all of these alternative latent variable models, only when all of the observed human rights variables are included in the estimation of the latent variable, does the trend line flatten out and become stagnant.

- See section 2.19 for a discussion of the relationship between a latent variable model estimates based on just the event-based variables, just the standards-based variables, and both (estimated using both the changing standard model and the constant standard model specification.

2.12

More generally, we show that alternative modeling choices have substantive consequences for answering questions about human rights improvement and for the development of human rights theory and relevant public policy. Any modeling strategy must assign weights to different types of evidence. We do not claim that a "correct" model should treat all of the human rights indicators similarly. Rather, we emphasize that the conclusion one reaches about the pattern of human rights improvement depends upon the specific weights assigned to two different types of evidence.

- Any measurement model must be identified so that the resulting parameters have meaning relative to one another. Cingranelli and Filippov propose a model that does not meet this criterion because the model is not identified with respect to time. This is a point I return to below and discuss at length in the main response to these authors. In brief, the model that Cingranelli and Filippov propose cannot be used for comparison of parameters from year to year.

2.13

The human rights subfield is engaged in a similar unproductive debate over whether human rights are improving or declining by focusing on different types of evidence. Scholars analyzing annual reports of commonplace repressive government practices such as torture and political imprisonment conclude that, in most countries, governments continue to violate human rights. On the other hand, scholars focusing on the decline in mass killings could conclude that human rights violations are becoming less common. Like the previous debate over the structure of power in US communities, neither position is right or wrong. The two types of evidence are conceptually different.

- See my original response to this claim in section 1.15.

2.14

The two most commonly used measures of human rights are the Political Terror Scale (PTS) and the CIRI Physical Integrity Index. According to Fariss these scores do not accurately record changes in human rights over time. Fariss presents the problem as a "changing standard of accountability". Human rights scores may be inconsistent over time, because: (a) human rights reports have gotten longer, and more information 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 (Clark and Sikkink 2013; Fariss 2014; Hafner-Burton and Ron 2009).

- See my original response to this claim in section 1.15.

2.15

For counterarguments and contrary evidence, see Richards (2016) and Haschke and Gibney (2017).

- Richards (2016) article is responding to Clark and Sikkink (2013). As Clark and Sikkink (2013) argue, coders may be influenced by larger quantities and greater quality of information when coding human rights reports. However, the argument made by Fariss (2014) builds on this idea but shifts the conceptual focus away from the coders and towards the producers of the human rights reports themselves: The US State Department and Amnesty International. Thus, Richards, like Cingranelli and Filippov, misunderstand the theory of the changing standard of accountability. It is not about academic coding procedures. Instead, it is about the monitoring agencies and the way the human rights reports themselves are produced. A coding procedure can be applied with complete fidelity and the changing standard of accountability can still influence what CIRI and PTS scores a particular case receives. The standards-based data are potentially biased not because the coding procedure is biased but because the reports themselves are produced by monitoring agencies that are changing the standards that they use in the process of documenting human rights abuse. Richards links the standard of accountability with the CIRI coding process in his article: "CIRI did change standards for certain variables over time as understandings of these human rights evolved in practice. However, when CIRI changed standards, the entire time series was recoded to include these new practices where reported" (491). To reiterate the coding procedures developed by the CIRI and PTS teams are not the standard of accountability adopted and developed to monitor human rights behaviors.

- Haschke and Gibney (2017) focus specifically on the coding process for one of the the two Political Terror Scale variables coded from the State Department reports. However, these authors only focus on the time period 1999-2015. These authors show that the average length of the State Department reports does not change over this period and that, though report length is associated with the categorical PTS score, it is not associated within-country variation. They use this evidence to infer that the PTS indicator based on these reports are not biased because of an increasing quantity of information. This inference is consistent with evidence for this period of time (1999-2015) presented in Fariss (2014). Specifically, Fariss (2014) shows in Figure 5 in the main article and and in Figure 11 in the supplementary appendix that the base-line probability of being coded at one of the five levels on the PTS scale is constant from approximately 2002 or 2003 through

2010 (updated to 2015 in this paper below). Prior to the period beginning in 2001 or 2002, there is evidence for a change in these thresholds parameters. Figure 18 displays this graphically.

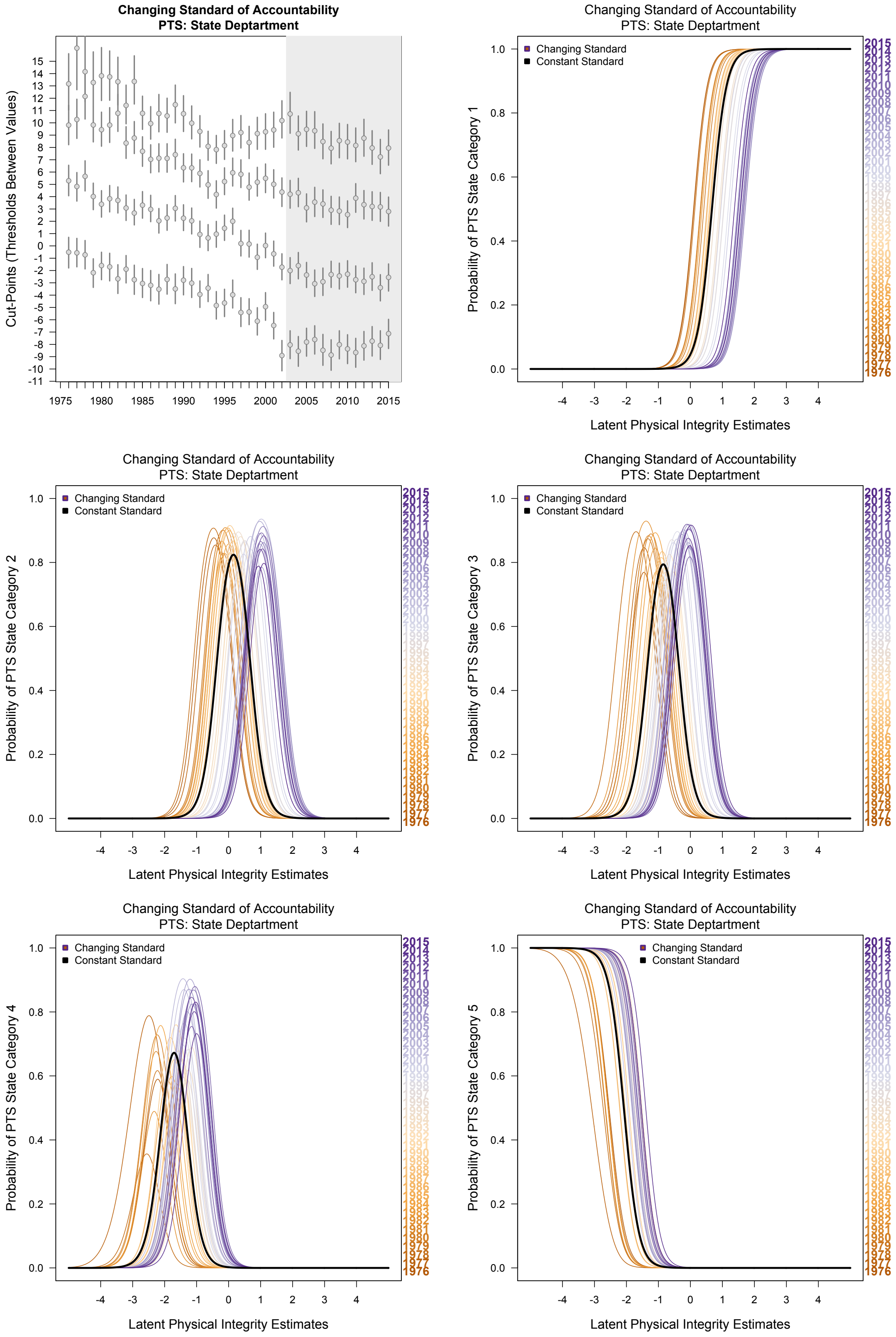

Figure 18: A decrease in the difficulty cut-points in the upper left panel translates directly into a change in the probability of being classified as a 1, 2, 3, 4, or 5 on the original PTS State Department Scale such that begin classified as 5 (e.g., frequent abuse) becomes more likely and 1 (e.g. no abuse) becomes less likely as a function of time. The threshold parameters and their corresponding probabilities stop decreasing for the period beginning in approximately 2002 through the most recent year of data in 2015.

2.16

Fariss suggested that possible biases in human rights data could be identified and corrected by estimating a latent index of human rights abuses. The assumption behind such an index is that, while the true level of human rights abuses is latent (i.e., unobserved), it is correlated with observable indicators of human rights. Various statistical techniques, ranging from factor analysis to IRT, would allow one to estimate a latent index based on observable values of several available indicators. Dynamic versions of IRT assume that criteria for recording the indicators could change over time. Fariss (2014) introduced a unique version of IRT to estimate latent human rights scores.

- See my original response in sections 1.17 and 1.16.

2.17

Combining two types of indicators, giving some weight to each type, should produce latent scores somewhere between the scores obtained when using each type separately. However, in Fariss's specification, that is not the case. When we replicated his analysis (using Fariss's computer code) we found that the reported upward trend in human rights depended almost entirely on the inclusion of the mass killings indicators. No indicators of lesser human rights violations were necessary. We replicated the analysis replacing the actual values of all indicators of lesser human rights violations with randomly generated data and obtained an identical trend (Figure 2 above). When we repeated Fariss's computations using only the five indicators of mass killing, again we obtained latent scores that show an improving trend (Figure 3 above) similar to the trend reported by Fariss (2014, 308). In contrast, when we replicated the analysis including only indicators of lesser human rights violations, there is no upward trend in the calculated latent scores (Figure 3).

- See my original response to this claim in section 1.22 and section 1.23.

2.18

A simple (bivariate) OLS regression shows that scores generated using only records of mass killing can explain 88% of the variation in Fariss's scores.

- Cingranelli and Fillipov argue that the event-based variables are responsible for most of the variation in the latent variable estimates. However, the squared correlation that Cingranelli and Fillipov present 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. For the period 1976-2015, the correlation between the latent variable point estimates based on just the event-based items (7 items) and the latent variable point estimates based on all of the items (16 items, changing standard of accountability model) is 0.74 (0.74 for the constant standard model). For the same period, the correlation between the latent variable point estimates based on just the standards-based items (9 items) and the latent variable point estimates based on all of the items (16 items, changing standard of accountability model) is 0.95 (0.99 for the constant standard model).

- For the period 1976-2015, the correlation between the latent variable point estimates based on just the event-based items (7 items) and the latent variable based point estimates on just the standards-based items (9 items) is 0.70.

- 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 1 because there are not standards-based variables as part of the model until 1976.

2.19

Figure 4 illustrates that the trends in Fariss's latent scores for many specific countries also are similar to the trends in latent scores produced only from records of mass killing.

- The evidence for the correlation coefficients (see above) suggests that the relationship between the latent variable models estimated with just the event-based variables, the standards-based variables, or all of these variables are very similar. The graphs presented below show these patterns visually.

- The latent variable model based on just the standards-based variables is much more closely related to the estimates from the full model from both the changing standard of accountability model and the constant standard of accountability model. All of the series tend to be closely related during the period from the late 1970s until approximately 1990.

- After 1989, there are only 3 event variables available for the event-based only series. The early period of 1949-1976, there are 6 event-based variables available, which explains why the credible intervals widen in the 1990s and 2000s for the event-based only series.

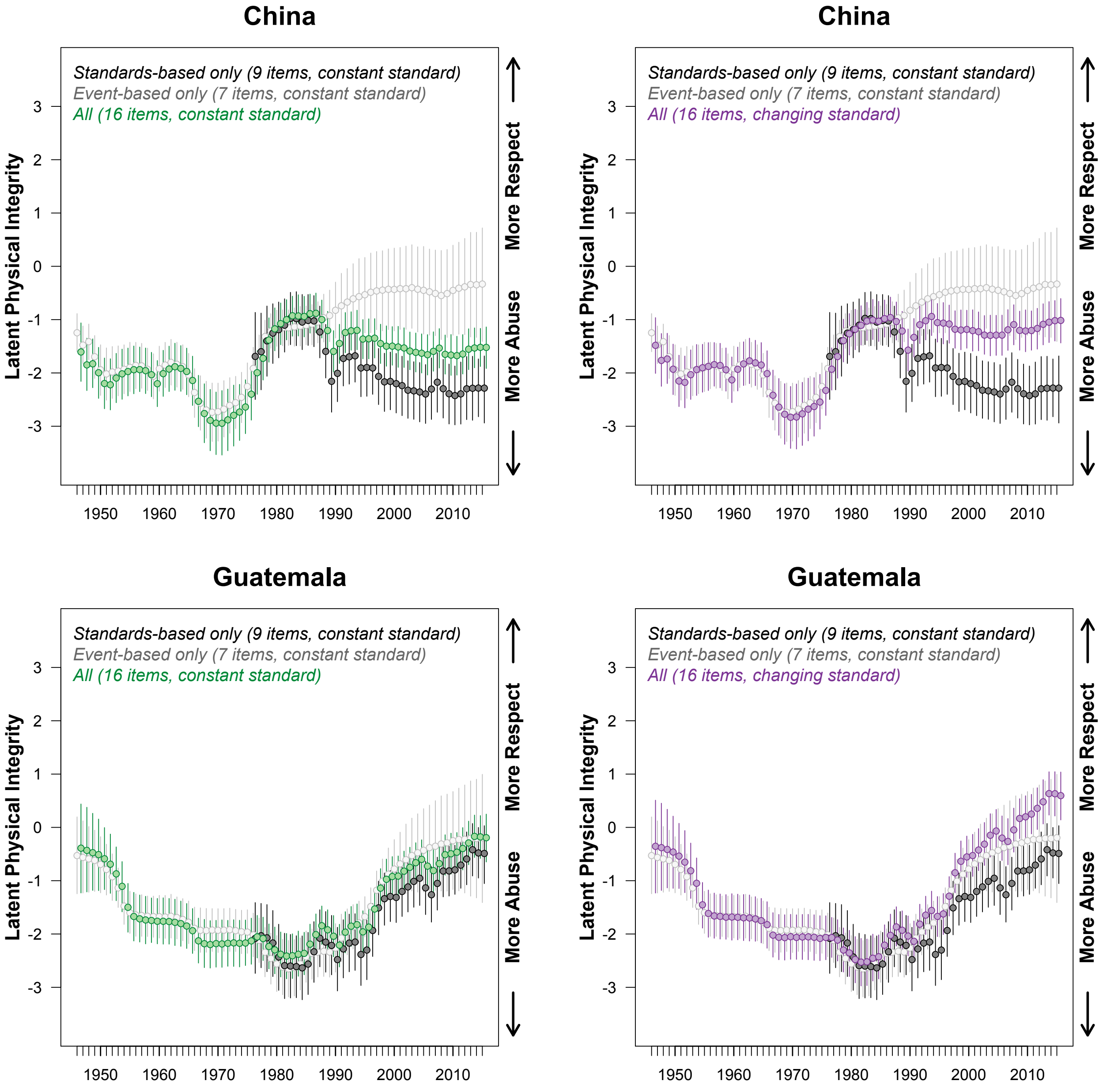

Figure 19: Latent variable estimates for two countries. The light grey trend is based on latent variable estimates using just the observed event-based variables. The dark grey trend is based on latent variable estimates using just the observed standards-based variables. The green or maroon trends are based on all of the variables using the constant standard model specification or the changing standard model specification respectively.

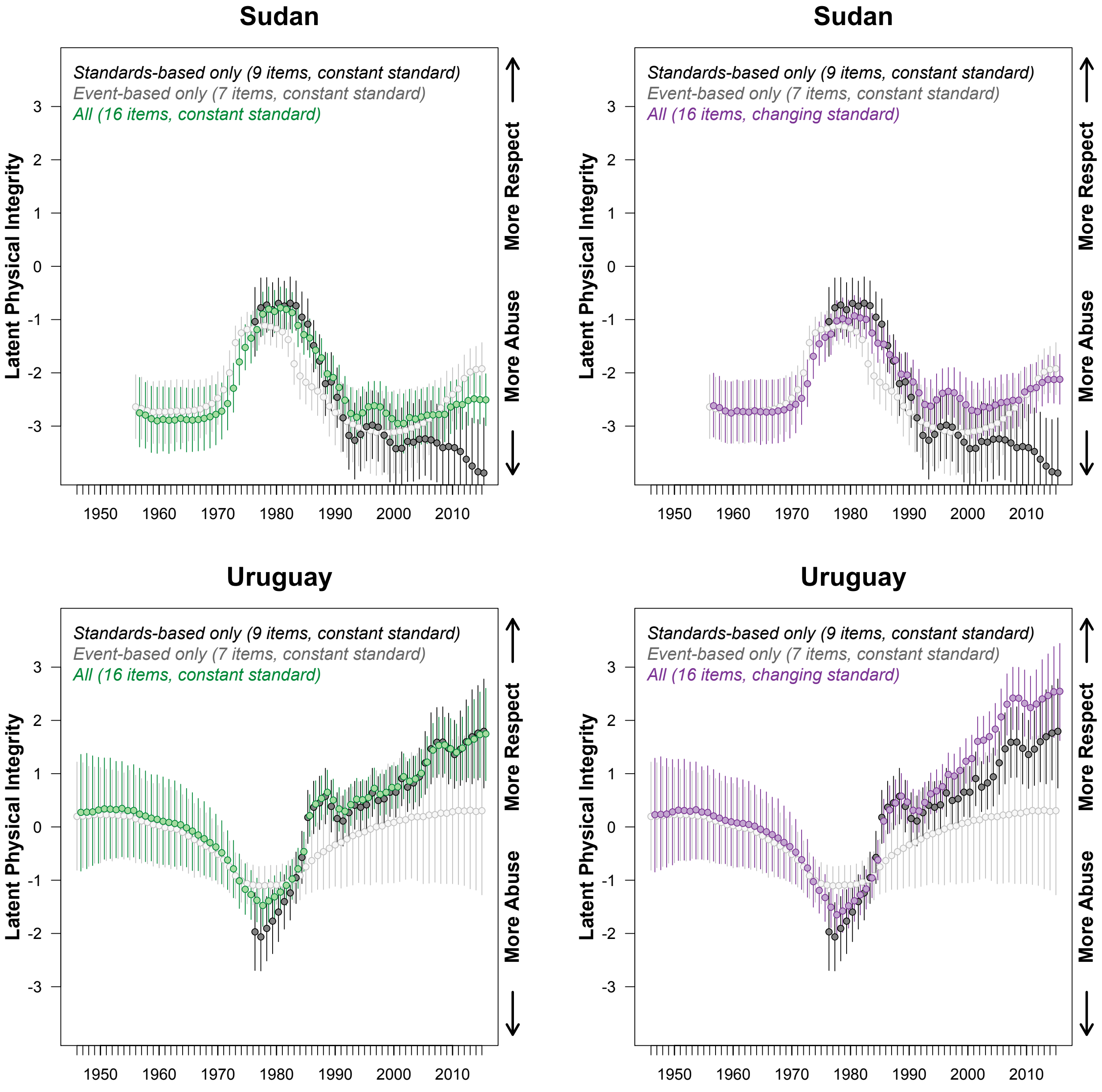

Figure 20: Latent variable estimates for two countries. The light grey trend is based on latent variable estimates using just the observed event-based variables. The dark grey trend is based on latent variable estimates using just the observed standards-based variables. The green or maroon trends are based on all of the variables using the constant standard model specification or the changing standard model specification respectively. The standards based variables are important for determining the within country changes over time for each country.

2.20

The customary way to use IRT is to treat all observable indicators similarly in their relationship with the latent variable. This is how dynamic latent IRT models were used previously (e.g., Martin and Quinn 2002; Schnakenberg and Fariss 2014; Wang et al 2013).

- A latent variable model must be identified with respect to the parameters the model should be able to estimate and compare.
- If some of the indicators are biased, then the relative placement of one unit relative to others will be biased too. The latent variable values are these placements.
- Again, differential item functions (DIF) allow the model to adjust the values of the latent variable given knowledge of the differences in the process by which the observed information arises. This is what the changing standard of accountability model does.
- For the constant standard model to be more consistent with reality, the human rights 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.
- I respond to this claim in much more detail in the main response and also in sections 1.26.

2.21

As applied to the debate in human rights, this approach would test the argument about the changing standards of accountability in human rights records by assuming that potentially all indicators are subject to such changes. Thus, they all could have variable intercepts. This approach leaves the possibility of a fixed intercept to be endogenously generated in the estimation.

- This model, as describe, is not identified with respect to time. It cannot test for the changing standard of accountability. It cannot account for any changes in the latent variable over time. This is because this model re-estimates the latent variable for the set of units in each year.
- I respond to this claim in much more detail in the main response and also in sections 1.27.

2.22

When Fariss combined the two types of data in a single estimator, he treated the two groups of indicators differently (Fariss 2014, 305-306). He set the mass killing indicators to follow a logistic regression with a fixed intercept (cut point) but he set the indicators of lesser human rights violations to follow an ordered logistic regression with variable intercepts for every year. Thus, the latent variable had to fit actual observations of the mass killing indicators without allowing a possible adjustment to the intercepts. With the indicators of lesser human rights. violations, on the contrary, it was much easier for the algorithm to fit the latent variable as there were several dozen additional parameters (time specific intercepts) that could also adjust. Consequently, variation in the mass killing indicators generated the improving trend of the latent variable. This is why the actual values of the indicators of lesser human rights violations did not matter much.

- See my original response to this claim in section 1.25.

2.23

In a private communication, Fariss acknowledged that fixing the intercepts for mass killing indicators was necessary to obtain the scores showing an improvement in human rights.

- Cingranelli and Filippov are referencing a public exchange that took place at the 2016 meeting of the Midwestern Political Science Association.

- This statement is incorrect. As Fariss (2014) documents and as I have discussed at length in the main response, all of the item difficulty parameters are fixed for the constant standard model. Allowing these parameters to vary over time for some items is possible. However, as the simulation analysis presented in the main response demonstrates, it is not possible for these parameters to vary over time for all of the items, which is the model preferred by Cingranello and Filippov.

- By allowing the the item difficulty parameters to vary over time for the observable variables derived from the Amnesty International and State Department reports to vary over time, an improving human rights trend emerges. This pattern is in comparison to the flat or stagnant trend that the constant standard model produces.

- For more details see my original response in section 1.6.

2.24

The estimates from a model where the intercepts for all items vary across time will produce latent variable estimates similar to those from a model where none of the intercepts vary. When we re-run Fariss's analysis allowing all indicators to have variable intercepts (in all other ways relying on Fariss's original computer code), we obtain the trend displayed in Figure 5, showing no human rights improvement. These results are robust to the choice of indicators included in the estimation from four individual CIRI components to all 13 available indicators. To summarize, Fariss found an improving trend in human rights only because his model did not allow the standards for recording mass killings to change.

- I respond to this claim in much more detail in the main response and also in sections 1.27.

2.25

The difference in the changes in the standards of accountability (S1 - S2) between the two types of records should be treated as an empirical question. It is likely that both lesser human rights violations and mass killing events are recorded more accurately now than in the past. There 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 becoming more or less consistent with one another.

- The changing standard of accountability is not about the coding rules used by the academic teams. See the main response to Cingranelli and Filippov.
- See my comments in section 1.10, and 1.15.

2.26

Fariss (2014, 301) assumed that instances of mass killings, genocides and political executions by oppressive regimes could "act as a consistent baseline" by which to compare the levels of variables measuring *lesser human rights violations. As we demonstrated above, the crucial assumption is that there has been no change in the "standards of accountability" in records of mass killings. We prefer an alternative, less restrictive assumption–that there has also been a change in the standards for recording mass killings as a starting point for thinking about ways to combine the two types of evidence.*

- The model that Cingranelli and Fillpov prefer is not capable of making comparisons over time. See the main response to Cingranelli and Filippov and section 1.27.

2.27

Though Fariss distinguishes between events-based and standards-based data, it is important to recognize that even mass killing indicators are standards-based. In order to record mass killings, scholars must make judgment calls that require coding rules determining such things as whether, and, under what circumstances, to include (a) relatively low death toll events (b) deaths due to interstate and civil war, and (c) killings by non-governmental actors such as paramilitaries.

- The theory of the changing standard of accountability is not about the coding procedures developed and implemented by the different human rights coding projects.
- I respond to this claim in much more detail in the main response and also in sections 1.32.

2.28

Harff and Gurr (1988, 365) developed relatively restrictive, explicit coding rules, only counting events in which "(a) many noncombatants were deliberately killed, (b) the death toll was high (in the thousands or more), and (c) the campaign was a protracted one." Like the PTS and CIRI coders, they relied on the annual reports issued by Amnesty International and the US State Department, among other sources, to identify their cases. Other mass killing scholars like Rummel (1994) applied less restrictive coding rules and recorded the greatest number of mass killings (Figure A3), even counting the United States as committing mass killings based on civilian wartime deaths in Korea and Vietnam.

- I respond to this claim in much more detail in the main response and also in sections 1.29, 1.30, 1.32.

2.29

Even if the standards for recording mass killings have been more constant, those extreme events should not be weighted so heavily that they become a proxy or substitute for more direct measures of lesser human rights violations. Mass killings and lesser forms of human rights violations may not even be indicators of the same underlying concept—respect for physical integrity human rights. The most general definition of the right to physical integrity is "freedom from 1) state-imposed deprivations of life, 2) physical harm at the hands of state agents, and 3) state-imposed detention" (Hill 2016). Lesser forms of physical integrity human rights abuses happen in almost every country every year. Mass killings are rare events, occurring mostly in failed and authoritarian states.

- Cingranelli and Filippov are suggesting that large scale killing events are a distinct repertoire of state-sanctioned repression in comparison to other forms of physical integrity abuses such political imprisonment, disappearances, and ill treatment and torture. However, these types of repressive practices are empirically related to state-sanctioned practices that are associated with extra-judicial killings and the large scale occurrence of killings as well. This conceptual understanding informs the primary 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 new argument made by Cingranelli and Filippov is actually similar to an argument made by McCormick and Mitchell (1997) two decades ago. Specifically, McCormick and Mitchell (1997) argue 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. Cingranelli and Richards (1999) argue against the conceptualization and use a statistic from Mokken (1971) to demonstrate the indicators of these four types of physical integrity violations scale together (see van Schuur, 2003, for additional information about Mokken scaling). This scaling result is supported by additional evidence about the relationship between all of the CIRI variables presented by Fariss and Schnakenberg (2014). Fariss (2014) discusses a multidimensional IRT model in the supplementary appendix of his article. There is no empirical support for a second dimension from the 13 indicators considered in that article.

2.30

More commonplace forms of human rights violations may be following one trend while the worst forms are following another (Gutierrez-Sanin and Wood 2017). Mass killings may have declined, because most governments have learned to selectively use the "lesser" forms of repression to achieve their objectives, while avoiding the worst and most notorious forms. Yet, Fariss's modeling specification forces the occurrence of the rare events of mass repression to strongly affect the human rights scores of all countries. Why should mass killing in Myanmar, for example, have any effect on the calculation of the scores of countries such as Denmark?

- If this is the pattern that is occurring, then human rights practices are, on average, improving.

- However, there is substantial evidence that certain forms of human rights abuses, what Cingranelli and Filippov now label as "lesser forms of abuse" were under-reported in earlier periods, when more egregious forms of abuse were prevalent. "Incidents 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" (Brysk, 1994, pg 681). This evidence is consistent with the highly transparent care of Sweden, with the CIRI human rights project codes as a case with ill treatment and torture occurrences (Eck and Fariss, 2018). See 1.14 and 1.30 for additional details about this substantively important deviant case.

- The latent variable model does not force the occurrence of rare events to "strongly affect the human rights scores of all countries". I respond to this portion of the above point in more detail in the next subsection.

- *Why should mass killing in Myanmar, for example, have any effect on the calculation of the scores of countries such as Denmark?* For example, let's consider the process of observations, categorization, and comparison in a world in which only these two cases exist. By necessity, a categorization scheme must be based on criteria that are mutually exclusive and jointly exhaustive. These criteria are necessary for the comparative method and have received considerable discussion in political science over the years (e.g., Kalleberg, 1966; Lijphart, 1971, 1975; Sartori, 1970). For example, the informational criteria used to create a binary categorization scheme about massive repression must be mutually exclusive so that evidence of the event leads to one coding and lack of evidence of the event leads to the other coding. The informational criteria is jointly exhaustive for this example because there are no other categorical options based on the definition and at least one of each event is observed in our simplified two-case world. To categorize cases in this way requires evidence about both types of events. This logic applies to more complex categorization schemes as well. Thus evidence from both Denmark, because there is currently no evidence that a massive repressive event occurred there, and evidence from Myanmar, because there is evidence that a massive repressive event occurred there, is necessary to jointly categorize them. The absence of evidence in the case of Denmark relative to the presence of evidence in Myanmar makes the categorization and therefore the comparison of this two-case example possible.

- The logic underlying the comparisons made from the latent variable estimates are the same. The latent variable model takes multiple categorical indicators and uses the categorized value of each case relative to the values of the other cases to estimate the latent variable, which is itself an estimate of the placement of each case relative to all of the other cases along a single dimension that spans the real line and that is governed by the normal density function. For any comparison, mutually exclusive and jointly exhaustive information from every case that is to be compared is necessary. The latent variable model is based on the probabilistic assessment of these placements.

- The latent variable model does not force the occurrence of rare events to necessarily be in a specific part of the latent space. Rather, the intercept or the cut points of the latent variable account for the relative frequency of each of the observed items, conditional on the value of the latent variable itself. The item weights, which are slope parameters from the logit or ordered logit link functions, are the parameters that determine, in conjunction with the intercept, the portion of the latent space that each of the country-year units will probabilistically occupy. Larger item weights — larger logit or ordered logit coefficients — represent increasing precision in the placement of the units along the latent trait when they take on a specific value of the observed item.

2.31

As a consequence of putting so much weight on rare mass killings, for many countries the trends in Fariss's scores are non-intuitive, especially those produced through backward data extrapolation and imputation for the 1949-1975 period. These trends would not stand up to the level of scrutiny applied by Clark and Sikkink (2013) who criticized PTS and CIRI scores for particular countries to illustrate the possibility of an information paradox in recording human rights violations.

- See section 1.13.

2.32

Particularly questionable are the trends in the scores between 1949 and 1975. No records of lesser human rights violation are available for this period. Moreover, for the majority of country-years between 1949 and 1976, the actual codes for all thirteen human rights indicators before extrapolation were ZERO. Yet all country-years were assigned different latent human rights scores.

- The claim that there is not coverage of "lesser forms" is no longer accurate. The latent variable estimates included in this response now make use of the negative sanctions variables from the World Handbook of Social and Political Indicators Taylor and Jodice (1983), which are available from 1948 through 1982.
- See section 1.13.

2.33

Despite research showing that democratic, economically developed countries have better human rights records (e.g., Poe and Tate 1994), authoritarian, less developed countries often rank higher than well-established democracies.

- See section 1.35.

2.34

Fariss places the United States in 1953 at the same level as North Korea. The US scores for the 1950s are well below the scores for South Africa, which, at the time, was building its apartheid regime. Between 1949 and 1970, the scores for the United States are significantly below the scores for Afghanistan and Soviet-satellite Mongolia.

- See section 1.14.
- Specifically, Fariss (2018*a*) comments on this point at great length in a published response to another critique by Cingranelli and Filippov. In particular, Fariss (2018*a*) discusses the deviant case of the United States in 1953. See also Eck and Fariss (2018) for a discussion of the deviant case of Sweden.

2.35

Fariss assumes that, because the worst forms of human rights violations have become less frequent, indicators of lesser violations of physical integrity such as torture should reflect this improving trend. If they do not, it is because records of those violations have been distorted by "changing standards of accountability." His model does not allow the records of mass killing also to be affected by changing standards of accountability. A model based on a less restrictive assumption, allowing both types of records to be affected by changing standards, however, leads to an estimation of no improvement in human rights latent scores.

- Any measurement model must be identified so that the resulting parameters have meaning relative to one another. Cingranelli and Filippov propose a model that does not meet this criterion because the model is not identified with respect to time. This is a point I discuss at length in the main response to these authors and in various portions of this appendix. In brief, the model that Cingranelli and Filippov propose is not identified with respect to time and cannot be used for comparison of parameters from year to year.

2.36

Computationally, the trend in Fariss's scores mirrors the trend in mass killing events. Emphasizing the decline in mass killings may be a researcher?s preference, but it makes Fariss's scores a poor source of inference about trends in physical integrity rights. The accumulation of research findings based on scores that are so heavily influenced by the declining frequency of mass killings will not contribute to better theories of why states why states engage in lesser forms of repression.

- See sections 1.2, 1.4, and 1.40.

2.37

Fariss was able to show that human rights are improving only because he relied on a unique functional form of dynamic IRT. However, his new method for calculating dynamic latent scores has not been properly evaluated. Its pitfalls and biases are unknown. The strengths and weaknesses of the previously used versions have been widely discussed in the literature. We can find no other example of using dynamic IRT where intercepts for some variables were allowed to vary but others were not.

- See sections 1.1 and 1.2.

2.38

Using a new model specification is not a reason for rejecting the results. However, we have identified serious problems. Replacing actual values with random data produced a similar improving trend. Random values, which have no trend, could not suffer from changing standards of accountability. Yet, our simulations show that Fariss's model specification would "correct" random values too.

- See section 1.7.

2.39

We conclude with some practical advice for scholars and policy analysts in the human rights subfield. Those who use Fariss's score should be aware that there is a strong built-in correlation between mass *killings and those scores. Policy evaluators should remember that the trends in Fariss's scores for capable and democratic countries are affected by frequencies of mass killing events in failed and authoritarian states. They also should realize that, as long as the frequency of mass killing does not return to pre-cold war levels, Fariss's model would produce latent scores with an improving trend in the future (Figure A.4). Thus, human rights will appear to improve subsequent to almost any policy intervention.*

- See section 1.8, 1.39, 1.40, and 1.41.

2.40

Also we remind scholars that latent scores should not be used as dependent variables in conventional regression analysis because doing so could produce inconsistent or severely biased estimates. Instead, when analyzing latent scores, it is necessary to use more advanced techniques such as simultaneous equation analysis, data simulation or multiple imputations (Bolck, Croon and Hagenaars 2004).

- See section 1.41.

2.41

Finally, we agree with Fariss that regardless of possible changes in standards of human rights recording, widely used indicators of human rights abuses (e.g., CIRI and PTS) are "useful for comparing state behaviors in the same year" (Fariss 2014, 303). Thus, a practical way to address concerns about the potential changing standards of accountability is to check the robustness of cross-national time-series results with cross-sectional estimations.

- See section 1.42.

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.

Bolck, Annabel, Marcel Croon and Jacques Hagenaars. 2004. "Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators." *Political Analysis* 12(1):3–27.

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

Caughey, Devin and Christopher Warshaw. 2016. "The Dynamics of State Policy Liberalism, 1936–2014." *American Journal of Political Science* 60(4):899–913.

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.

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.

Clark, Ann Marie and Kathryn Sikkink. 2016. "Response to David Richards." *Human Rights Quarterly* 38(2):493–496.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

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* .

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.

Fariss, Christopher J and Scott A Tyson. 2018. "How Data comes to Be." *working paper* .

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, MA: Cambridge University Press.

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):171–183.

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.

Hare, Christopher, David A. Armstrong II, Ryan Bakker, Royce Carroll and Keith T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

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.

Imai, Kosuke, James Lo and Jonathan Olmsted. 2017. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* forthcoming.

Kalleberg, Arthur L. 1966. "The Logic of Comparison: A Methodological Note on the Comparative Study of Political Systems." *World Politics* 19(1):69–82.

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.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65(3):682–693.

Lijphart, Arend. 1975. "The Comparable-Cases Strategy in Comparative Research." *Comparative Political Studies* 8(2):158–177.

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):134–153.

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.

Mislevy, Robert. 1991. "Randomization-based inference about latent variables from complex samples." *Psychometrika* 56(2):177–196.

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: .

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):853–872.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press.

Poole, Keith T. and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35(1):228–278.

Poole, Keith T. and Howard Rosenthal. 1997. *A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

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." *https://dx.doi.org/10.2139/ssrn.2828703* .

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

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

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4):1033–1053.

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

Seawright, Jason and John Gerring. 2008. "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61(2):294–308.

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.

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):139–163.