# Supplementary Appendix:
# Yes, Human Rights Practices Are Improving Over Time

Christopher J. Fariss*

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

# Contents

# Introduction to Appendix

This supplementary appendix 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*b*). This appendix provides additional discussion in relationship to three critiques: one published in the *British Journal of Political Science* and two versions of a newly published paper in *American Political Science Review*, all written by Cingranelli and Filippov (2018*a*; 2018*b*). In each case, these authors critique of the use of a particular specification of a latent variable model to understand patterns of human rights over time (Fariss, 2014*b*). Additional evidence can be found in the main response to this critique and an additional response to the first critique (Fariss, 2018*a*,*b*).

The two versions of the new critique contain many factually incorrect statements. Because of this, I have also written a line-by-line rebuttal. Some of the material contained in this appendix is adapted from these line by line responses, which I wrote after reading each new version of this article (Cingranelli and Filippov, 2018*a*). Please note that the appendix that accompanies Fariss (2014*b*) is available with the replication material associated with that paper (Fariss, 2014*a*) and on my professional website here: `cfariss.com`. Overall, Cingranelli and Filippov have provided little evidence to support their argument that the that human rights are not improving over time and that the standard of accountability is not changing, explicitly made in Cingranelli and Filippov (2018*b*) and implied by their argument and evidence in Cingranelli and Filippov (2018*a*).

Below, Appendix A provides variable definitions, source references and citations for each of the 16 observed human rights variables. Appendix B presents trends in the observed human rights variables over time. Appendix C shows variation between the latent variable estimates and the original observed human rights variables. Appendix D discusses the parameters from the latent variable models. Appendix E discusses the difference between face validity and concurrent validity with respect to the relationship between the theoretical concept and the latent variable specification and evidence from the latent variable as it corresponds to the status of particular cases or groups of cases. Appendix F discusses what an extrapolation is and how this is a distinct form of estimation in comparison to the latent variable model. Appendix G reviews the comparative method and how to make comparisons between diverse cases such

as Denmark and Myanmar. Appendix H presents information from several deviant or unexpected cases such as Sweden today and the United States in 1953. Appendix I presents the latent variable trends over time for democracies and non-democracies. Appendix J reviews suggestions for using the latent variable estimates in applied research.

# A Observed Human Rights Variables Descriptions and Citations

I have incorporated three additional indicators in the latent variable model of human rights. Two of these new indicators are binary event-based variables. One codes state-sponsored mass killing since the end of WWII (Ulfelder and Valentino, 2008), which is available from 1946 to 2015. The other event-based variable is a negative sanctions variable taken from the World Handbook of Social and Political Indicators data (Taylor and Jodice, 1983), which is available from 1948 to 1982. The third, is a new standards-based variable from the Political Terror Scale, which codes annual human rights reports from the monitoring organization Human Rights Watch, which is available from 2013 to 2015.

Table 1: Standards-Based Repression Data Sources (9 items)

| **Dataset Name** and Variable Description | **Dataset Citation** and Primary Source Information |
|---|---|
| **CIRI Physical Integrity Data, 1981-2011**<br>1. political imprisonment (ordered scale, 0-2)<br>2. torture (ordered scale, 0-2)<br>3. extrajudicial killing (ordered scale, 0-2)<br>4. disappearance (ordered scale, 0-2) | Cingranelli, Richards and Clay (2015)<br>Amnesty International Reports[1] and<br>State Department Reports[2]<br>*Information in Amnesty reports takes precedence over information in State Department reports* |
| **Hathaway Torture Data, 1985-1999**<br>5. torture (ordered scale, 1-5) | Hathaway (2002)<br>State Department Reports[1] |
| **Ill-Treatment and Torture (ITT), 1995-2005**<br>6. torture (ordered scale, 0-5) | Conrad and Moore (2011),<br>Conrad, Haglund and Moore (2013),<br>Amnesty International (2006)<br>Annual Reports[1], press releases[1],<br>and Urgent Action Alerts[1] |
| **PTS Political Terror Scale, 1976-2015**<br>7. Amnesty International scale (ordered scale, 1-5)<br>8. State Department scale (ordered scale, 1-5)<br>9. Human Rights Watch scale (ordered scale, 1-5)*<br>* Human Rights Watch scale (2013-2015) | Gibney et al. (2017),<br>Gibney and Dalton (1996)<br>Amnesty International Reports[1]<br>State Department Reports[1]<br>Human Rights Watch Reports[1] |

*1. Primary Source; 2. Secondary Source*

Table 2: Event-Based Repression Data Sources (7 items)

| **Dataset Name**<br>and Variable Description | **Dataset Citation**<br>and Primary Source Information |
|---|---|
| **Ulfelder and Valentino Dataset, 1946-2015**<br>10. massive killing<br>(categorized as 1 if event occurred; 0 otherwise) | Ulfelder and Valentino (2008)<br>historical sources (see article references)[1] |
| **Harff and Gurr Dataset, 1946-1988**<br>11. massive repressive events<br>(categorized as 1 if event occurred; 0 otherwise) | Harff and Gurr (1988)<br>historical sources (see article references)[1] |
| **Political Instability Task Force (PITF), 1956-2010**<br>12. genocide and politicide<br>(1 if country-year experienced event; 0 otherwise) | Harff (2003), Marshall, Gurr and Harff (2009)<br>historical sources (see article references)[1]<br>State Department Reports[2]<br>Amnesty International Reports[2] |
| **Rummel Dataset, 1949-1987**<br>13. genocide and democide<br>(categorized as 1 if event occurred; 0 otherwise)<br>(3 death count estimates: best, low, high) | Rummel (1994, 1995),<br>Wayman and Tago (2010)<br>New York Times[1], New International Yearbook[2],<br>Facts on File[2], Britannica Book of the Year[2],<br>Deadline Data on World Affairs[2],<br>Kessing's Contemporary Archives[2] |
| **UCDP One-sided Violence Dataset, 1989-2015**<br>14. government killing (event count estimate)<br>(1 if country-year experienced event 0 otherwise)<br>(3 death count estimates: best, low, high) | Eck and Hultman (2007), Sundberg (2009)<br>Reuters News[1], BBC World Monitoring[1]<br>Agence France Presse[1], Xinhua News Agency[1],<br>Dow Jones International News[1], UN Reports[2],<br>Amnesty International Reports[2],<br>Human Rights Watch Reports[2],<br>local level NGO reports (not listed)[2] |
| **World Handbook of Political and Social Indicators WHPSI, 1948-1982**<br>15. political executions (event count estimate)<br>16. negative sanctions (event count estimate)<br>(categorized as 1 if event occurred; 0 otherwise) | Taylor and Jodice (1983)<br>New York Times[1], Middle East Journal[2],<br>Asian Recorder[2], Archiv der Genenwart[2]<br>African Diary[2], Current Digest of Soviet Press[2] |

*1. Primary Source; 2. Secondary Source*

# B Observed Human Rights Variables Over Time

In Figure 1 and Figure 2 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.

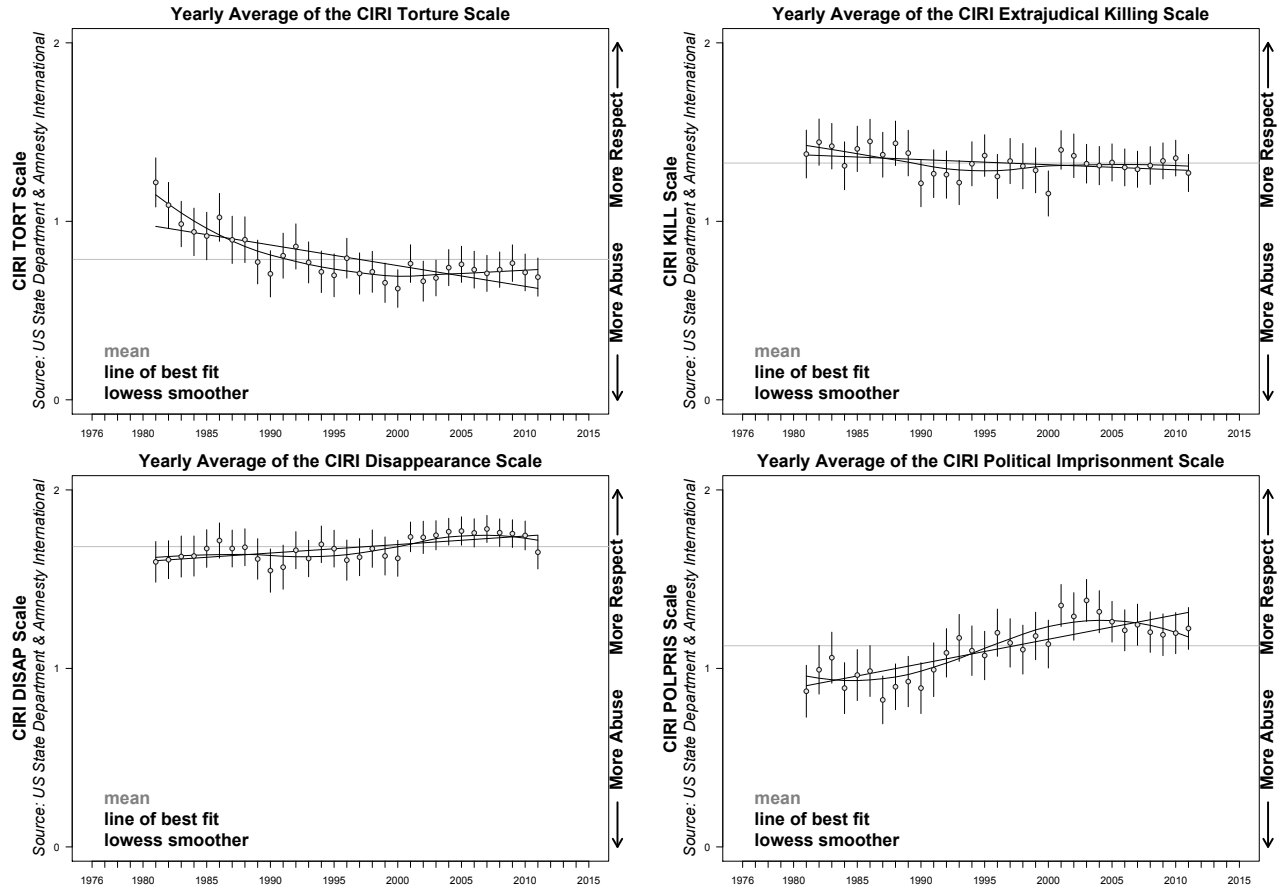


Figure 1: Trends in standards-based human rights variables over time. Note that all of these variables except the ITT and CIRI political imprisonment variables enter the changing standard of accountability model with varying item difficulty parameters (one for each year). The ITT and CIRI political imprisonment variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

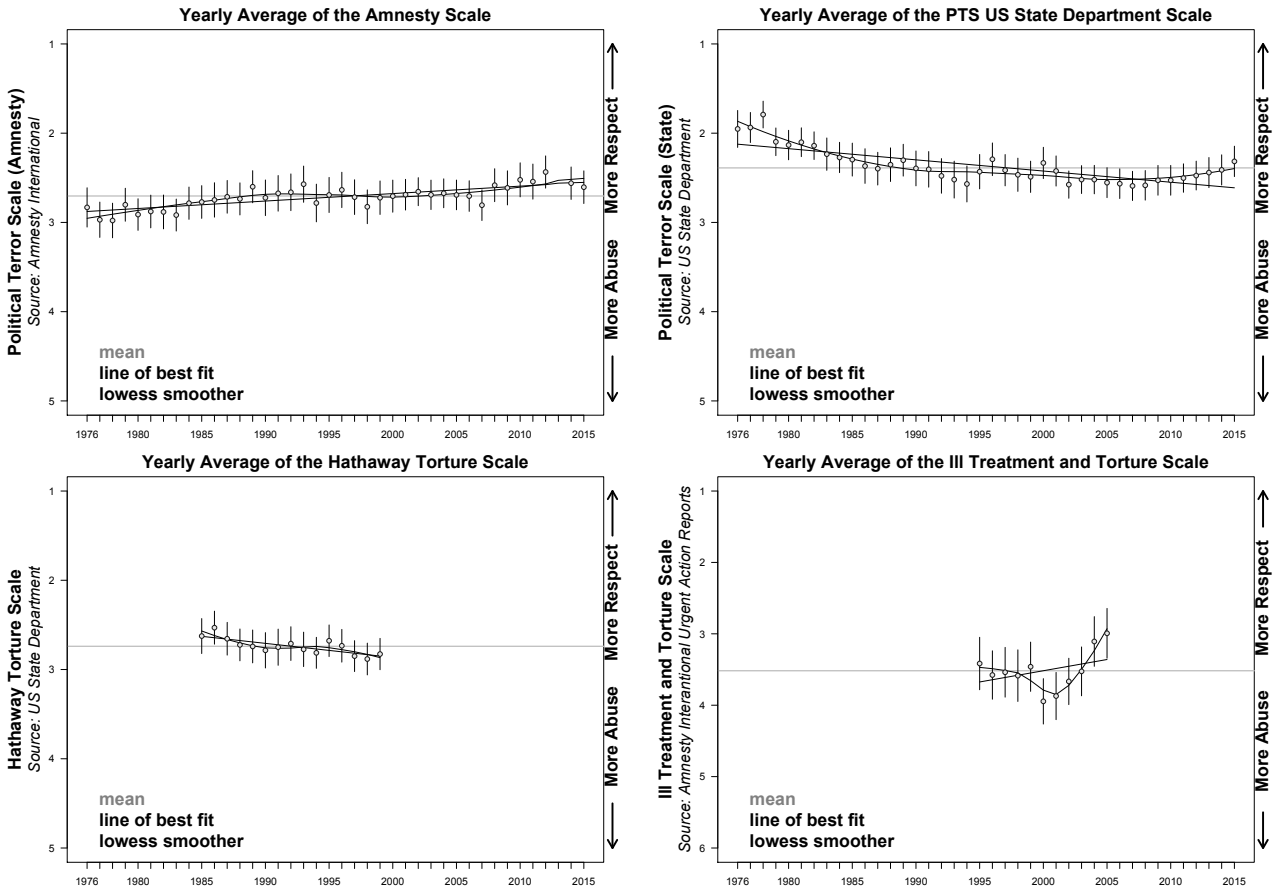


Figure 2: Trends in standards-based human rights variables over time. Note that all of these variables except the ITT and CIRI political imprisonment variables enter the changing standard of accountability model with varying item difficulty parameters (one for each year). The ITT and CIRI political imprisonment variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

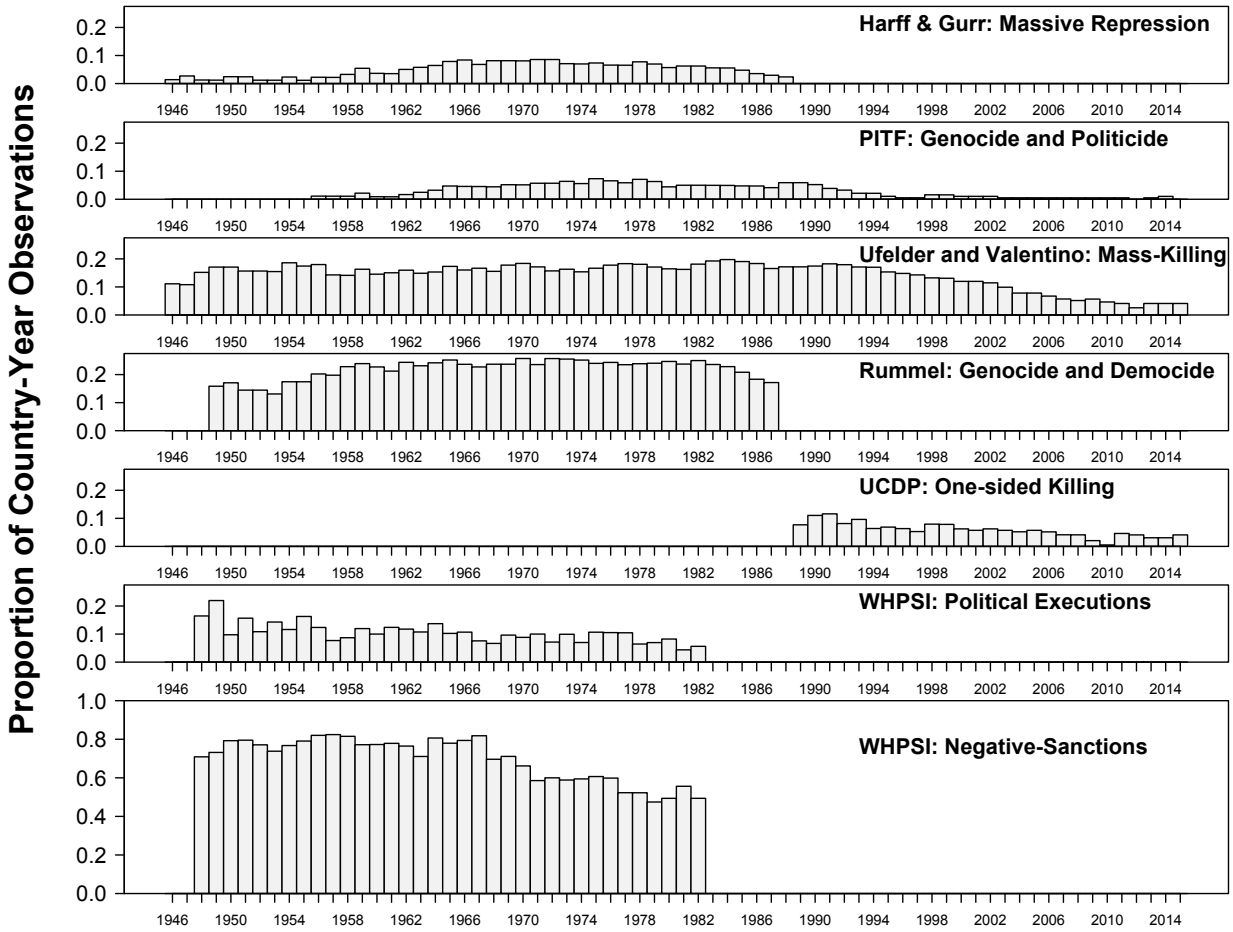


Figure 3: Trends in events-based human rights variables over time. Note that these variables enter the changing standard of accountability model with constant item difficulty parameters (one for all years).

# C Variation between the Latent Variable Estimates and the 16 Observed Human Rights Variables

Figure 4, 5, 6, and 7 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.

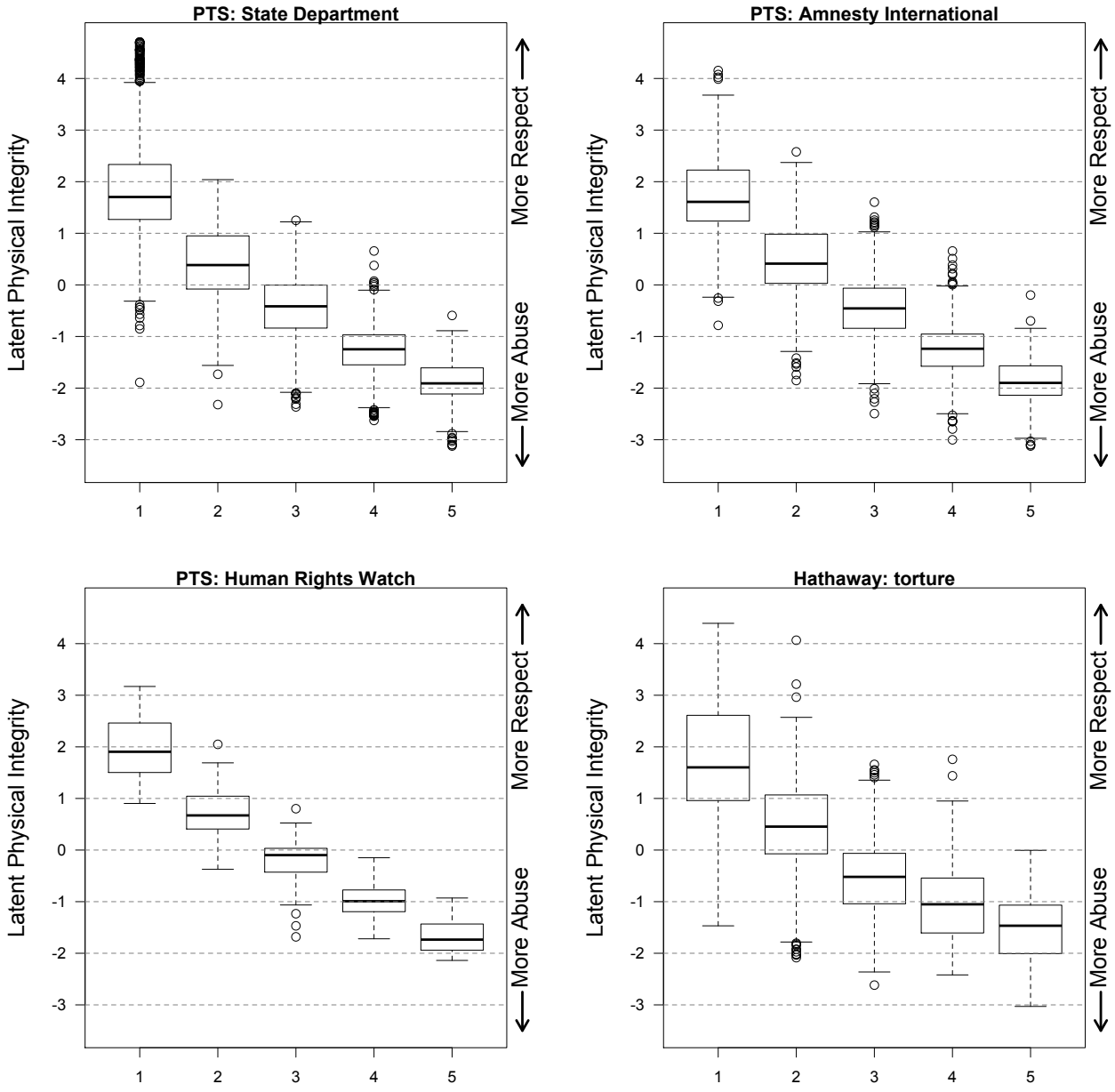


Figure 4: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014*b*) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human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

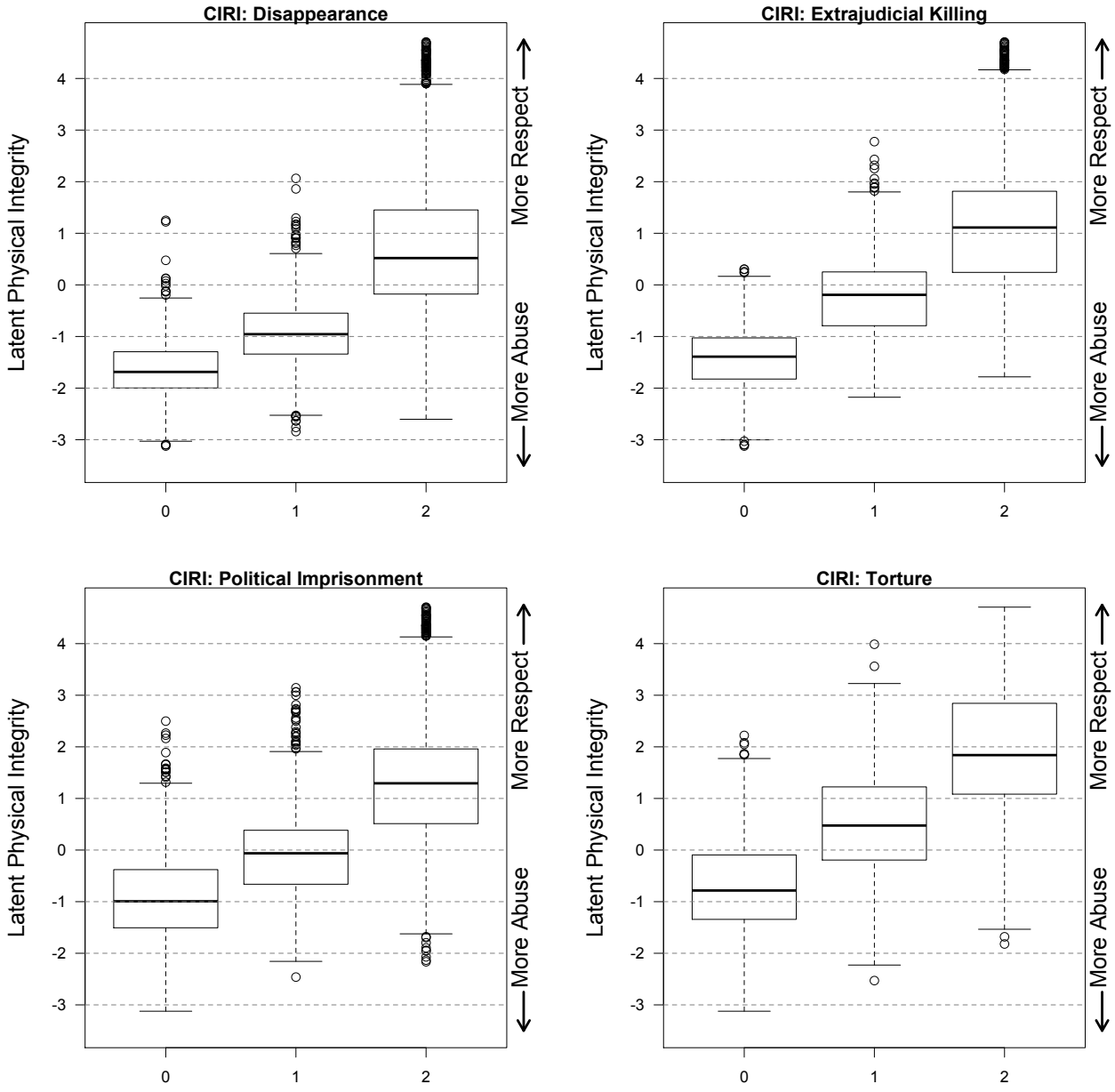


Figure 5: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014*b*) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014*b*) provides detailed documentation for each in the supplementary appendix that accompanies that article.

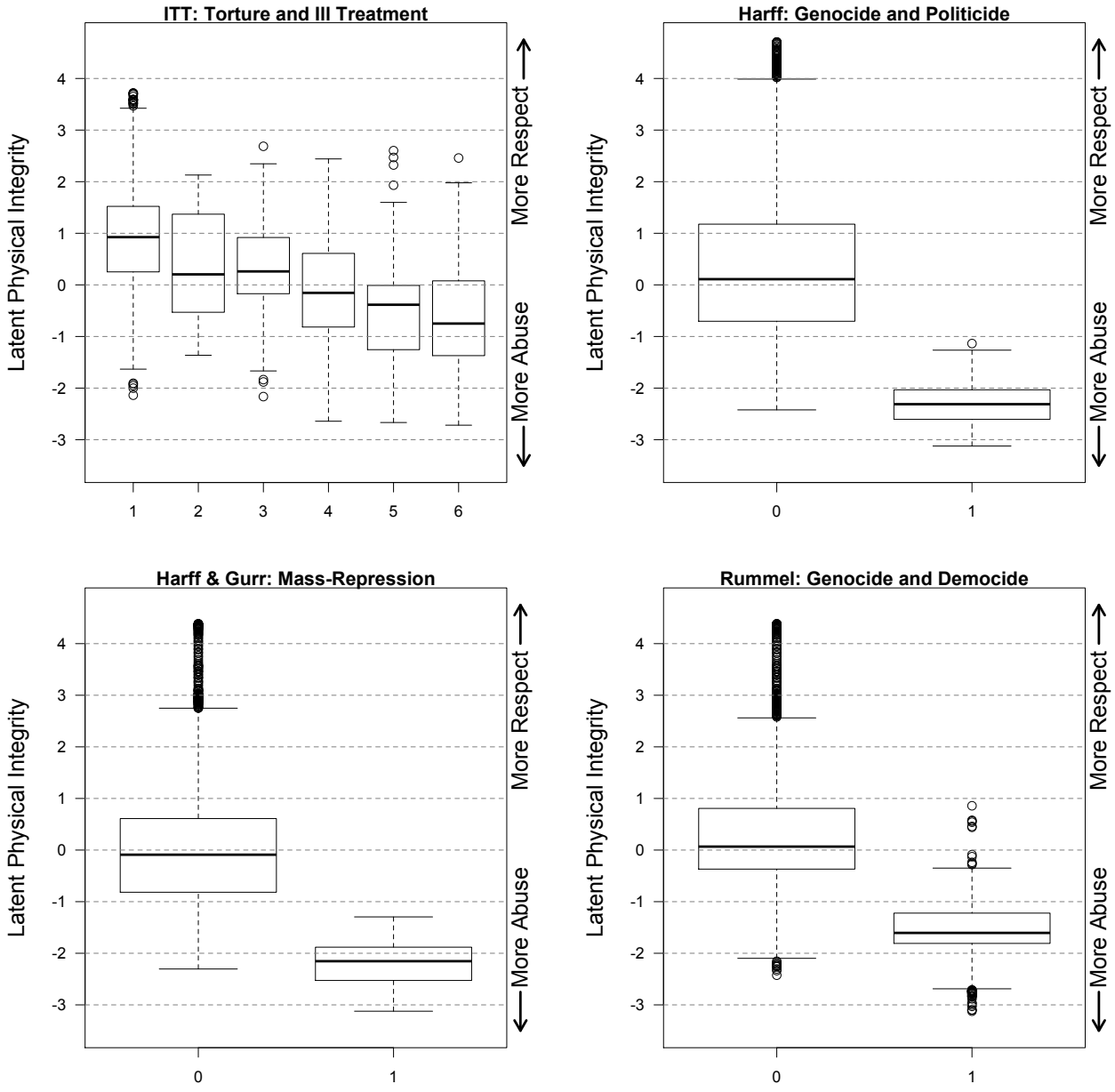


Figure 6: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014*b*) reflects variation in each of the 16 observed variables now included in the model.Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014*b*) provides detailed documentation for each in the supplementary appendix that accompanies that article.

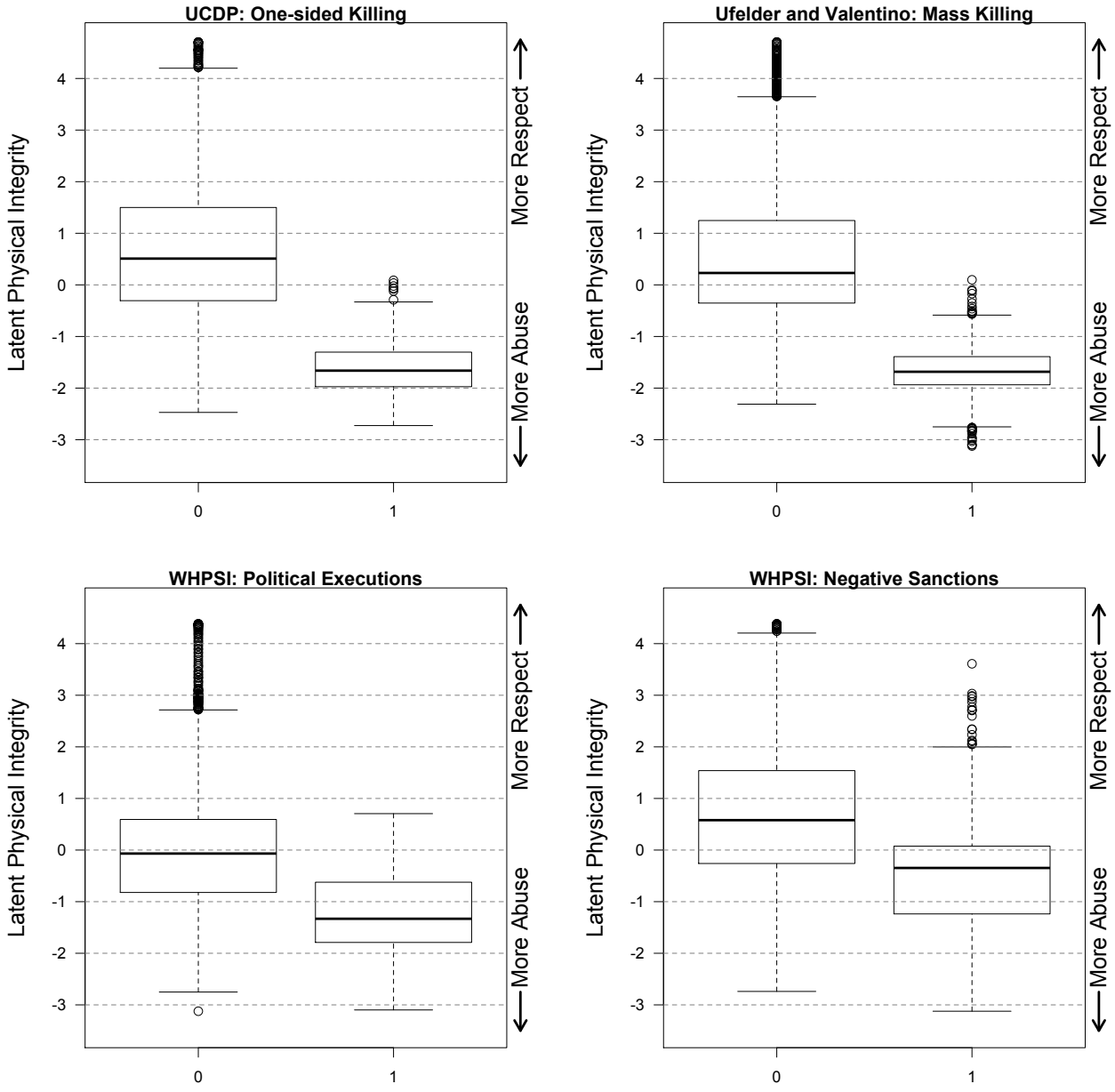


Figure 7: Variation in the latent scores from the changing standard of accountability models presented in Fariss (2014*b*) reflects variation in each of the 16 observed variables now included in the model. Values along the x-axis correspond to the categorical coding values for the specific observed human rights variable. (Fariss, 2014*b*) provides detailed documentation for each in the supplementary appendix that accompanies that article.

# D Item Weights and Item Values

In their abstract, Cingranelli and Filippov (2018*a*) state that Fariss (2014*b*) relies on "stringent assumptions" to argue he "heavily weighted rare incidents of mass killings such as genocide" (pg. 1). This is not true. The latent variable models in Fariss (2014*b*) incorporates the event-based variables in exactly the same way between the constant standard model and the changing standard model. The standards-based variables are treated differently, but it is not through the item-weights, but rather through their item difficulty parameters. The term item-weights is usually used to describe the item discrimination parameters in the IRT models, which are analogous to slope parameters in a logit or ordered-logit function, whereas the item difficulty parameters are analogous to intercepts for the binary human rights data.[1] Fariss (2014*b*) treats the standards-based variables differently in accordance with the theory of a changing standard of accountability, which I discuss in great detail in the main manuscript and in Fariss (2014*b*).

Here, I provide more details about 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 logistic function over the range of the possible latent variable estimates. 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 probabilities for each item value change over the range of values of the latent trait (Figure 8 shows this visually for several simulated examples). The item-discrimination parameters (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 (above or below a certain point) 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

[1] In the case of ordered categorical human rights data, these are instead sets of cut-points.

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. As the size of $\beta$ increases, the point along the latent dimension at which country0year units fall above or below becomes increasingly precise or clear relative to the other observed items.

Figure 8 illustrates the relationship between the different model parameters, the latent trait itself and the item-discrimination parameter and item-difficulty parameter for three binary items. The negative ratio between the the item-discrimination parameter and the item-difficulty parameter are governed by the range of possible value for the latent trait which is the standard normal density function. Figure 8 shows item response curves. From top to bottom the item discrimination parameters for $\beta$ are set to 1, 3, and 9 respectively. For the $\beta = 1$ case, the intercepts happen to fall on the same scale as the latent trait. The inflection point is the point at which a unit falling on the latent variable interval has a 0.50 probability of a 1 for the particular binary item. The inflection point for each is $-\frac{\alpha}{\beta}$. To scale the item-discrimination parameters with the logit function, much larger item difficulty parameters are necessary. From left to right $-\frac{\alpha}{\beta}$, the inflection point for each of the three items, is 2, 0, and -2 respectively. Note though that the intercepts themselves increase as *beta* increases. So the higher the precision of the item weight, the larger the intercept value needs to be to scale the inflection point along the range of the latent trait.

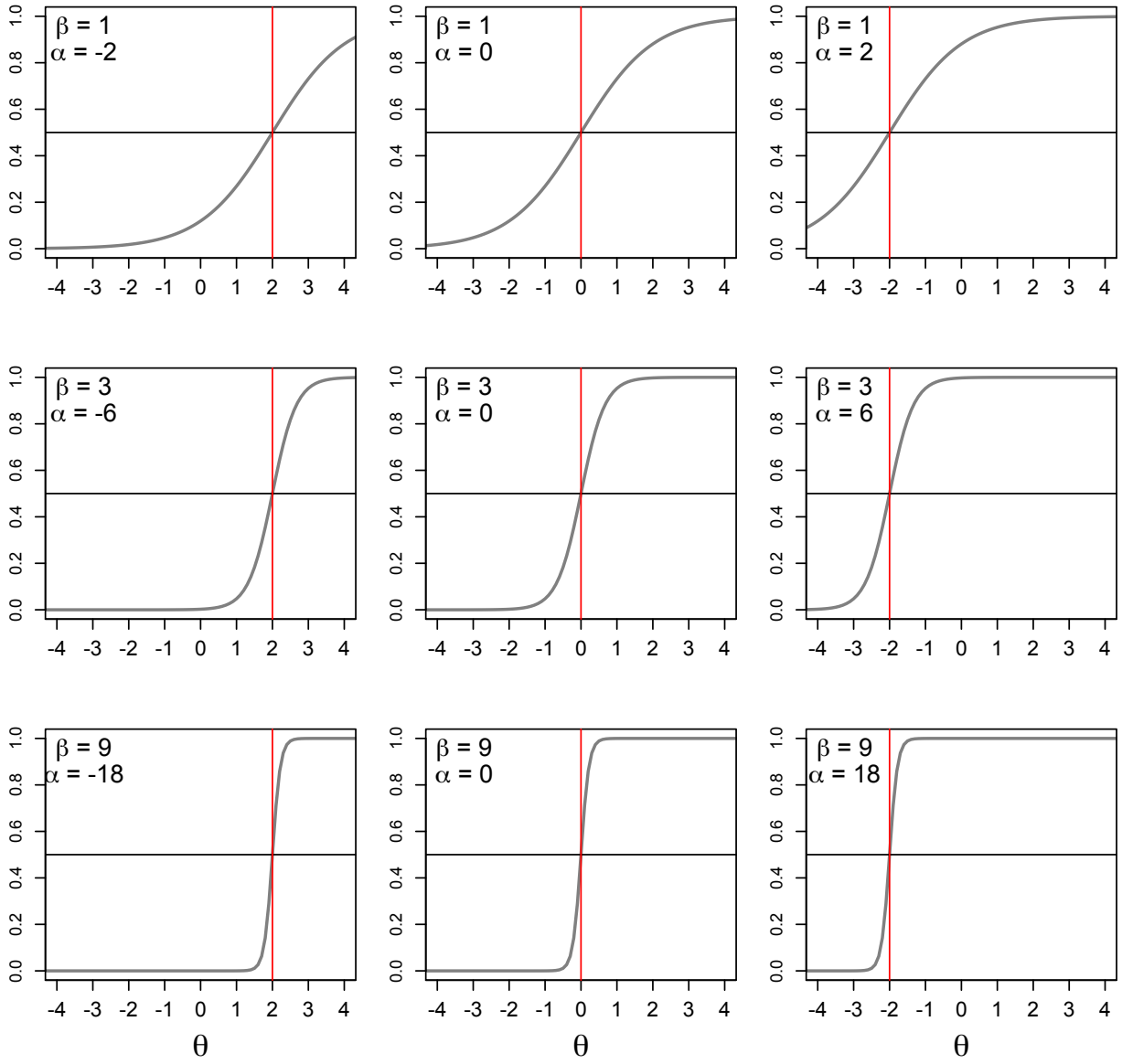


Figure 8: Item-response curves for three binary items. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along the latent variable that the units will probabilistically occupy. Larger item weights — 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 negative ratio between the the item-discrimination parameter and the item-difficulty parameter are governed by the range of possible value for the latent trait which is the standard normal density function. The red line is the inflection point, which is the point at which a unit falling on the latent variable interval has a 0.50 probability of a 1 for the particular binary item.

Figure 9, 10, 11, 12, 13, 14, 15, 16, and 17 display this visually below for the event-based and standards-based variables. What the figures show, is that 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. The curves inflection points reside in practically the same spot along the latent trait between the two models for each item. What this means, is that the models are treating the event-based variables the same. The standards-based variables are treated differently however because item-difficulty parameters are estimated for each year for some of these variables. These trends corroborate that the inference that not all of the standards-based variables change in the same way over time. The variables that change the most each year are the CIRI torture variable and the PTS variable based on the State Department report. These two variables have the strongest trends over time. In all cases the constant standard probability is similar to the probability in the earlier years for each of these variables.

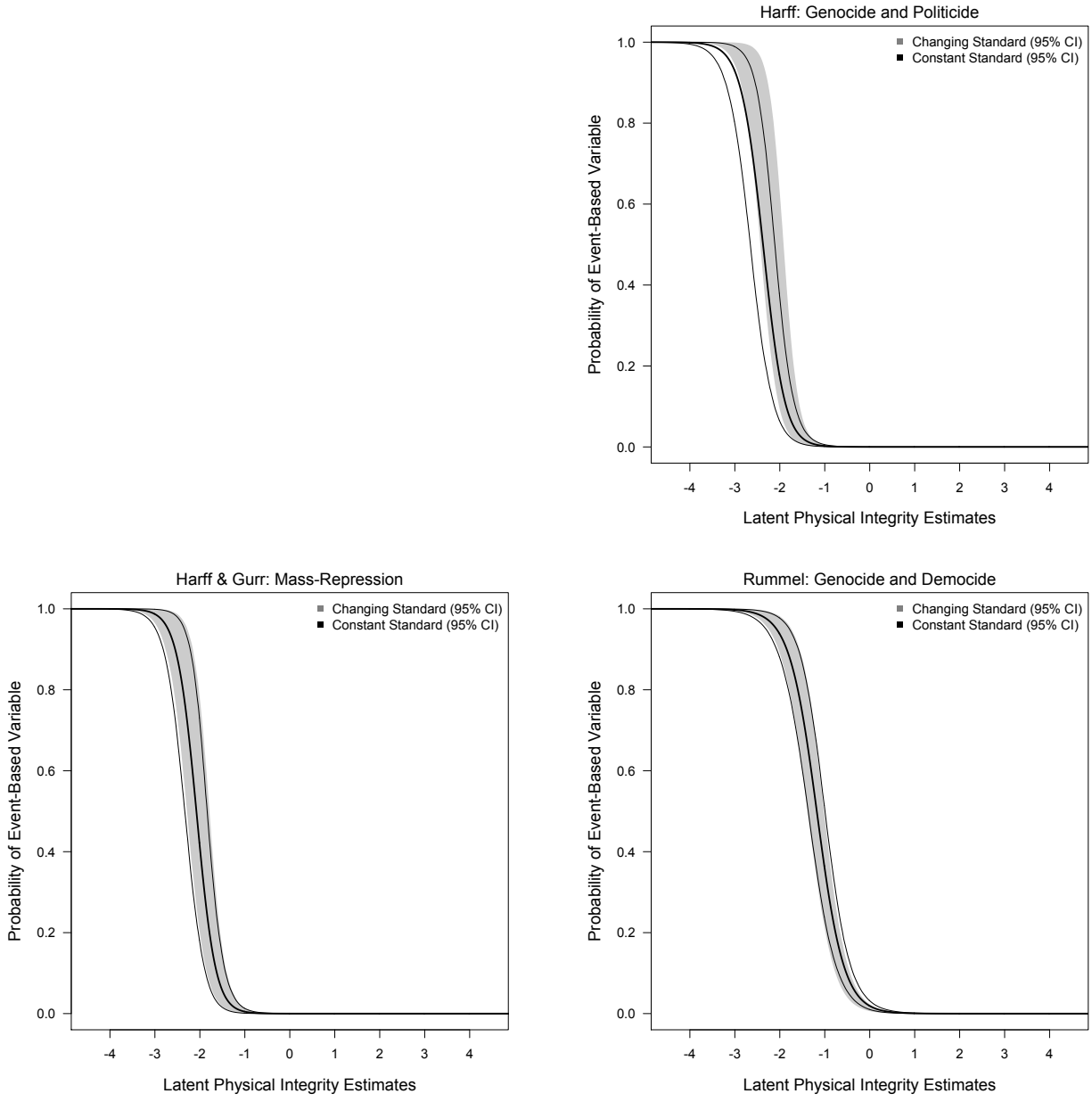


Figure 9: Item-response curves for the event-based variables. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along the latent variable that each of the country-year units will probabilistically 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 (grey) and constant standard models (black).

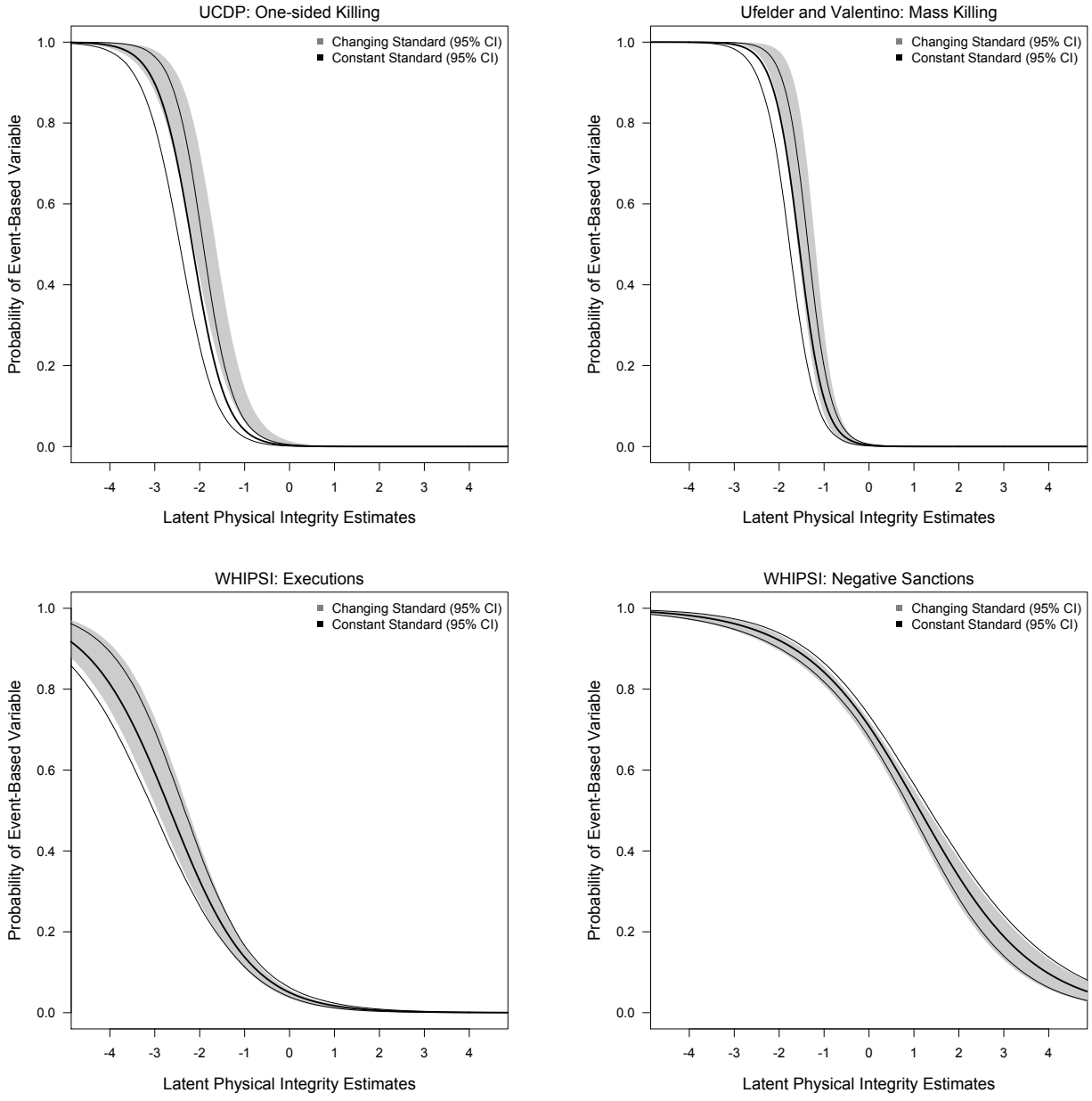


Figure 10: Item-response curves for the event-based variables. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the intercepts), the position along the latent variable that each of the country-year units will probabilistically 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 (grey) and constant standard models (black).

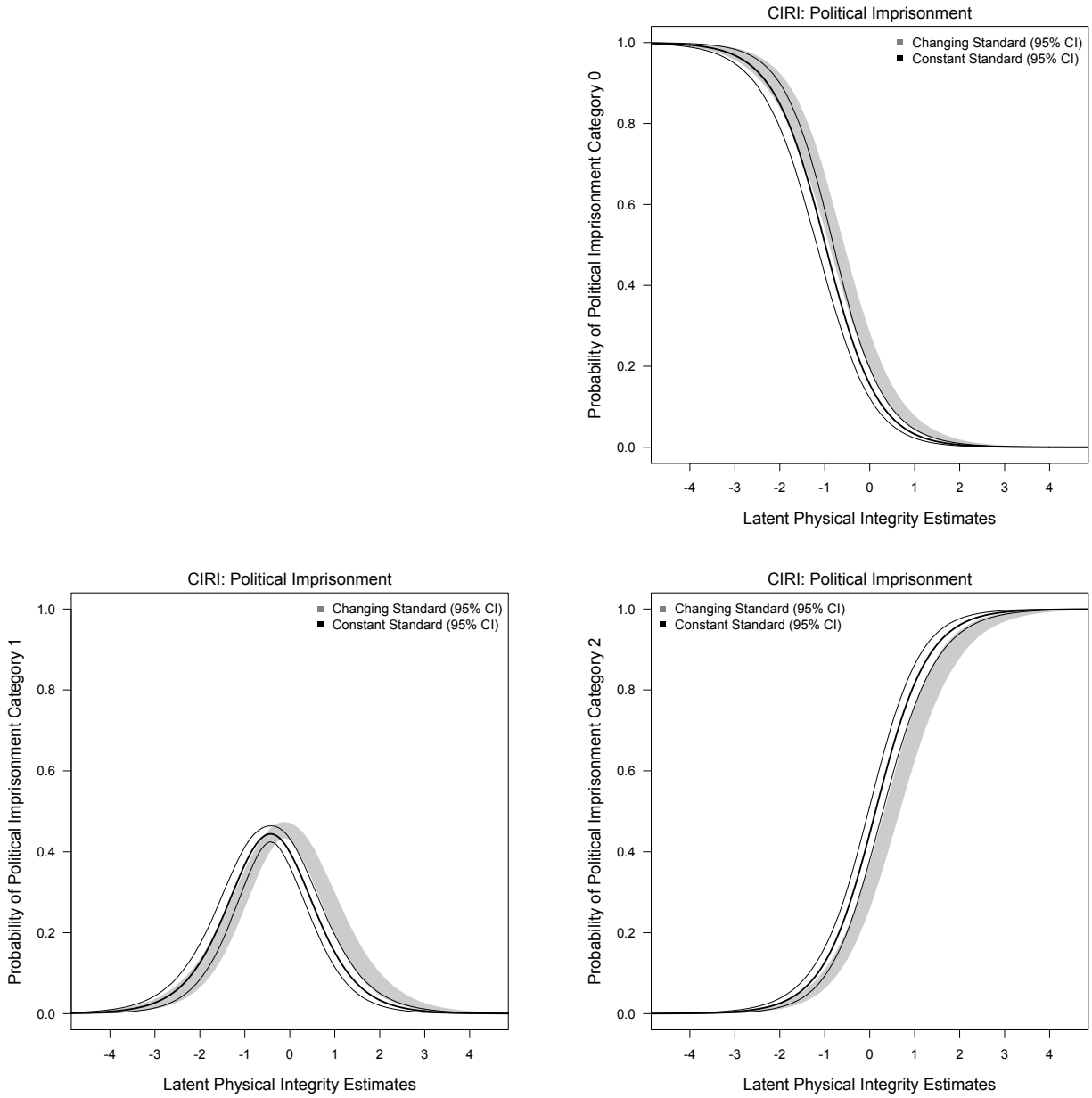


Figure 11: Item-response curves for each level of the CIRI: political imprisonment variable. Category 0 is the worst value on the scale, category 1 is the middle value on the scale, and category 2 is the best value on the scale. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the cut-points), the position along the latent variable that each of the country-year units will probabilistically occupy. The item-difficulty parameter is the same for each probability cure. Larger item weights — larger 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. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (grey) and constant standard models (black).

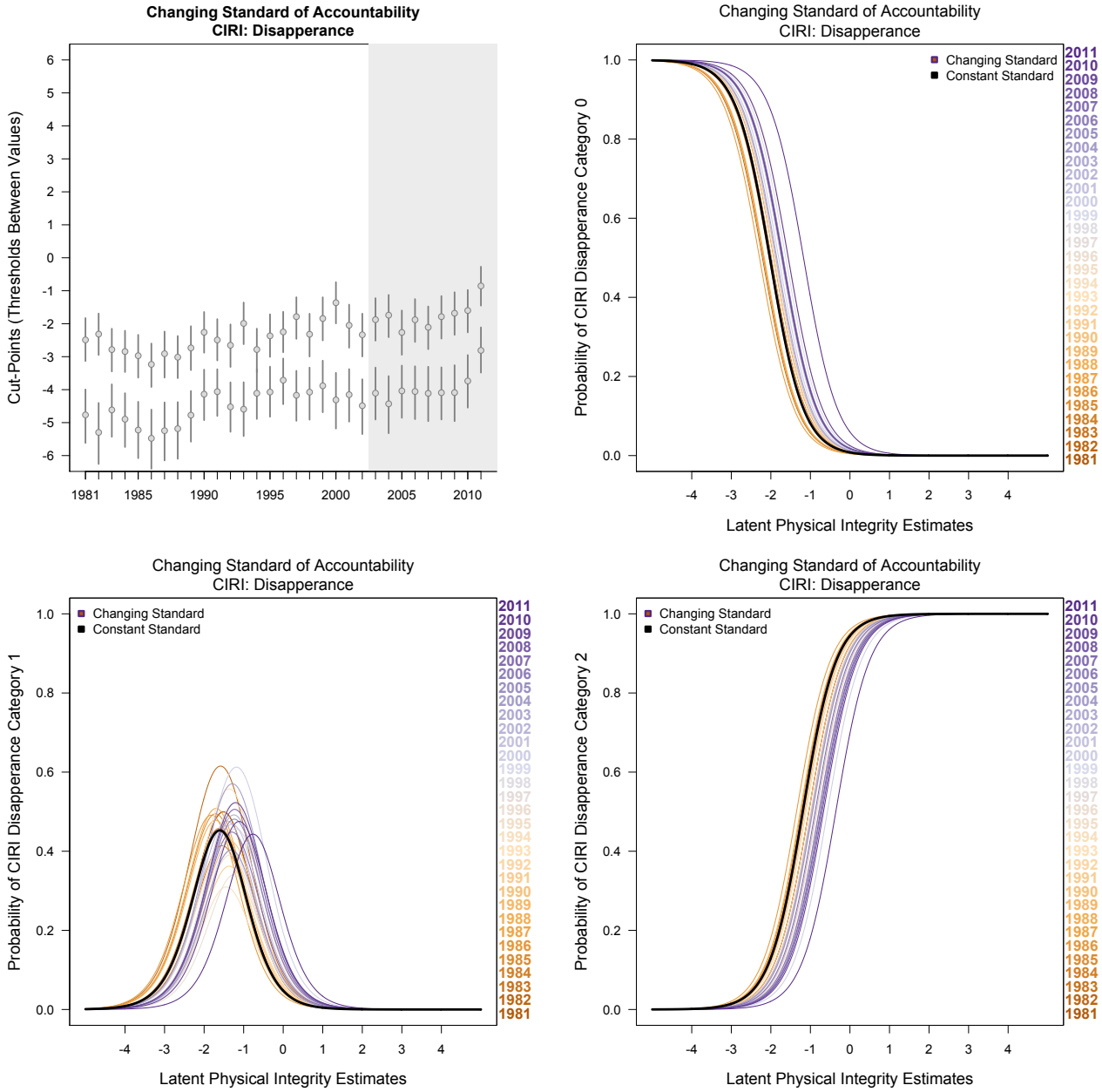


Figure 12: Item-response curves for each level of the CIRI: disaperances variable. Category 0 is the worst value on the scale, category 1 is the middle value on the scale, and category 2 is the best value on the scale. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the cut-points), the position along the latent variable that each of the country-year units will probabilistically occupy. The item-difficulty parameter is the same for each probability cure. Larger item weights — larger 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. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (grey) and constant standard models (black).

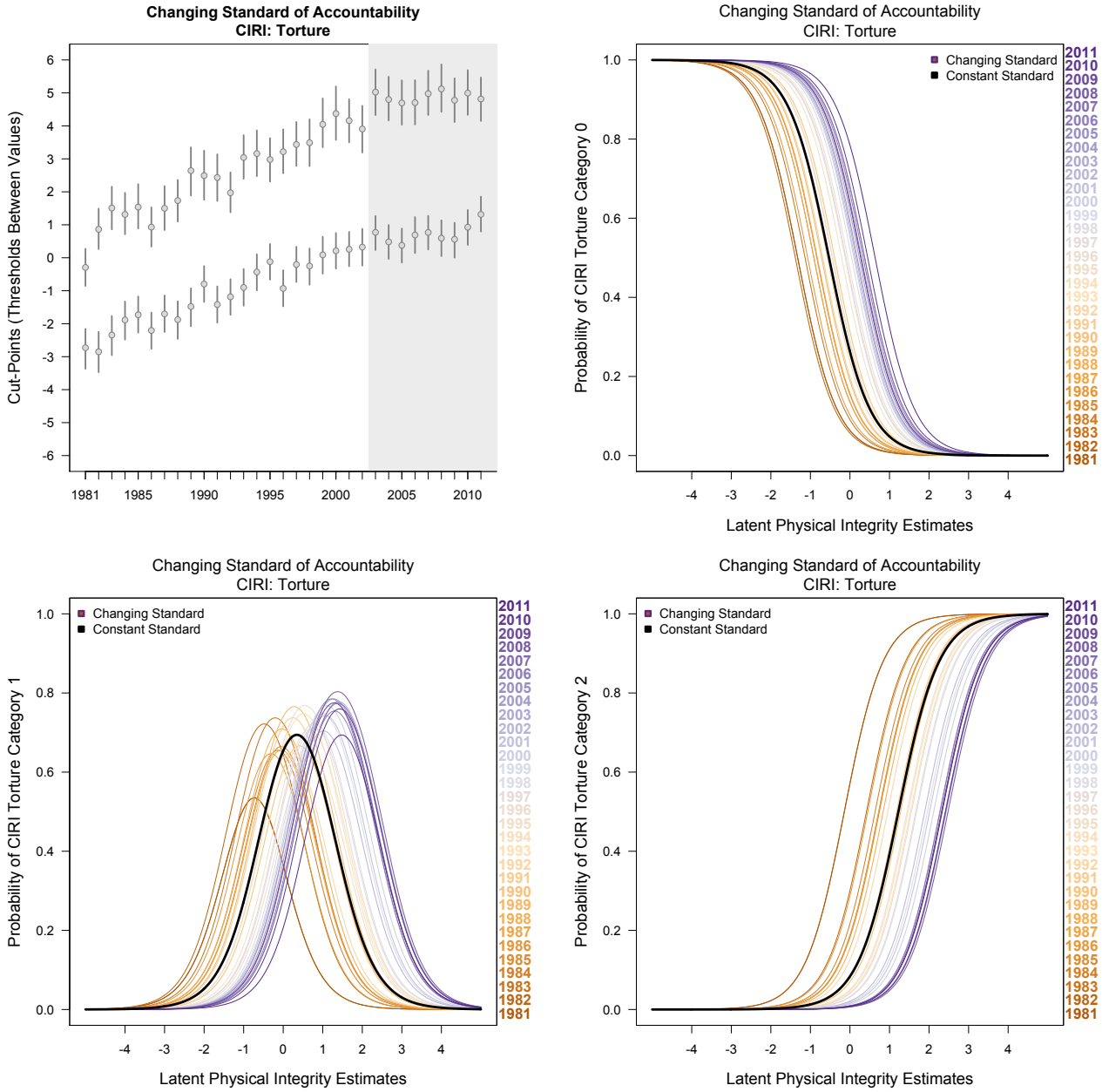


Figure 13: Item-response curves for each level of the CIRI: torture variable. Category 0 is the worst value on the scale, category 1 is the middle value on the scale, and category 2 is the best value on the scale. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the cut-points), the position along the latent variable that each of the country-year units will probabilistically occupy. The item-difficulty parameter is the same for each probability cure. Larger item weights — larger 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. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (grey) and constant standard models (black).

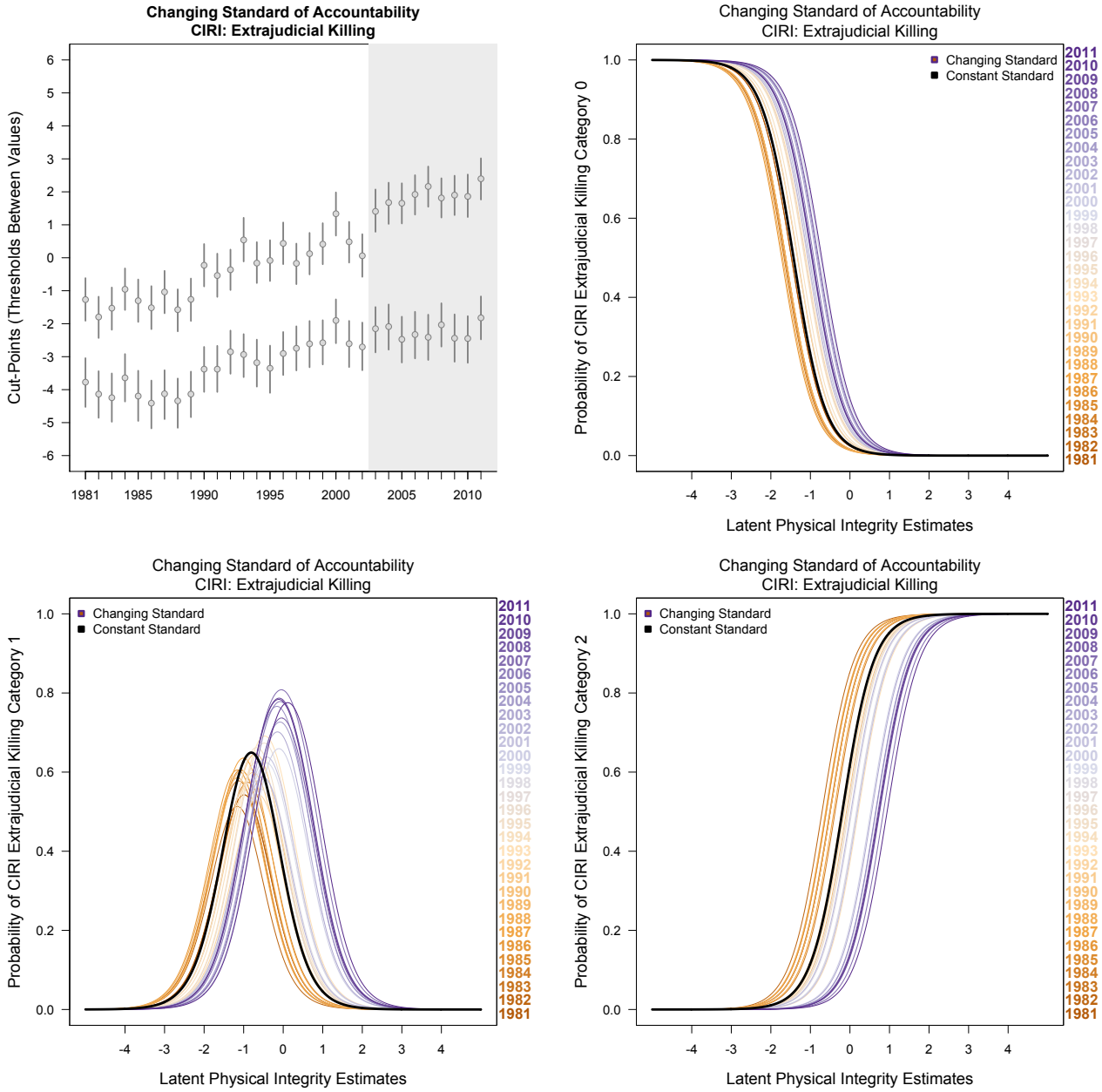


Figure 14: Item-response curves for each level of the CIRI: extrajudicial killing variable. Category 0 is the worst value on the scale, category 1 is the middle value on the scale, and category 2 is the best value on the scale. The item-weights (i.e., the slopes) are the parameters that determine, in conjunction with the difficulty parameters (i.e., the cut-points), the position along the latent variable that each of the country-year units will probabilistically occupy. The item-difficulty parameter is the same for each probability cure. Larger item weights — larger 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. The probability distribution for each of the event-based items are probabilistically quite similar when compared between the changing standard (grey) and constant standard models (black).

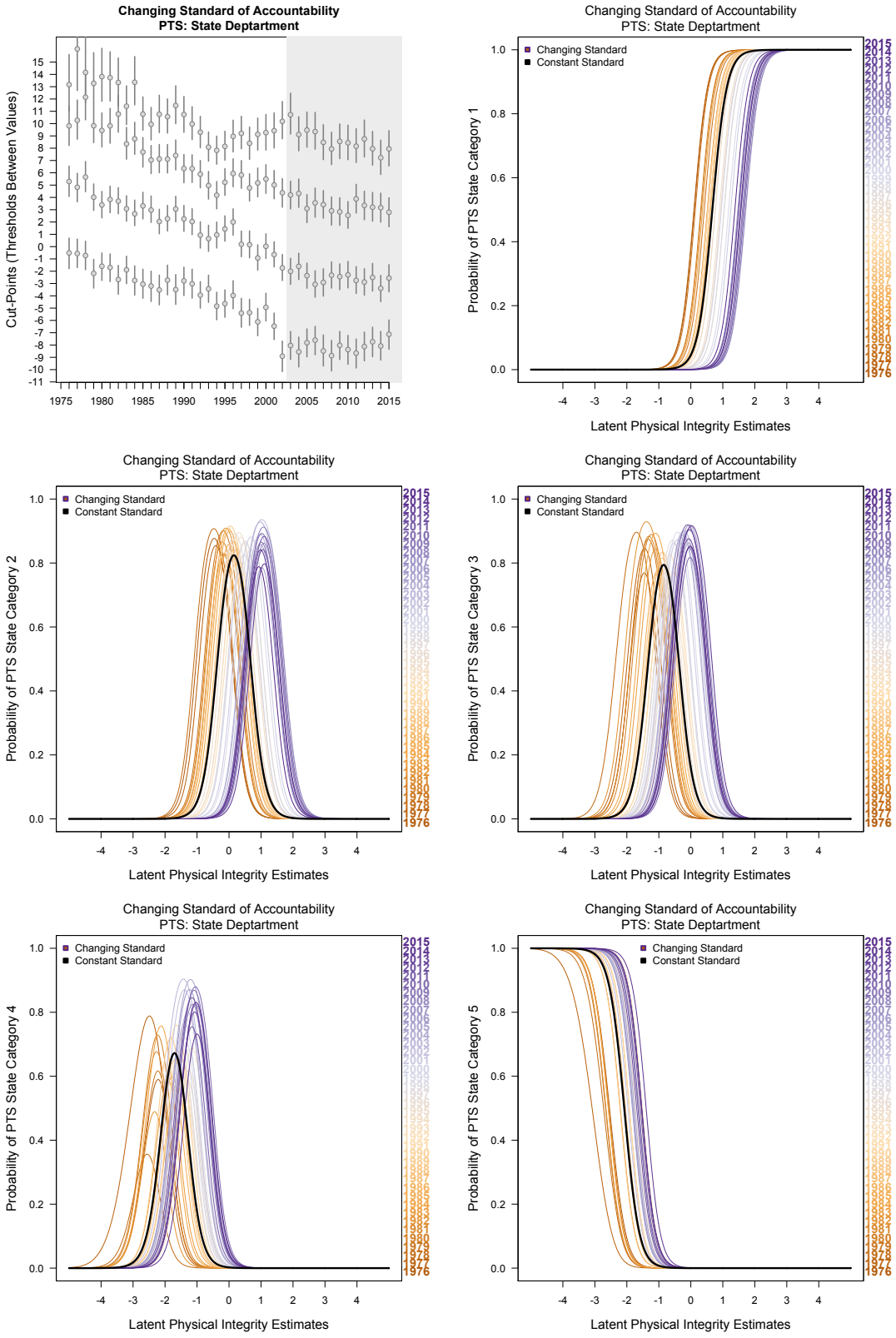


Figure 15: A decrease in the difficulty cut-points in the upper left panel translates directly into a change in the probability of being classified as a 1, 2, 3, 4, or 5 on the original PTS State Department Scale such that begin classified as 5 (e.g., frequent abuse) becomes more likely and 1 (e.g. no 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.

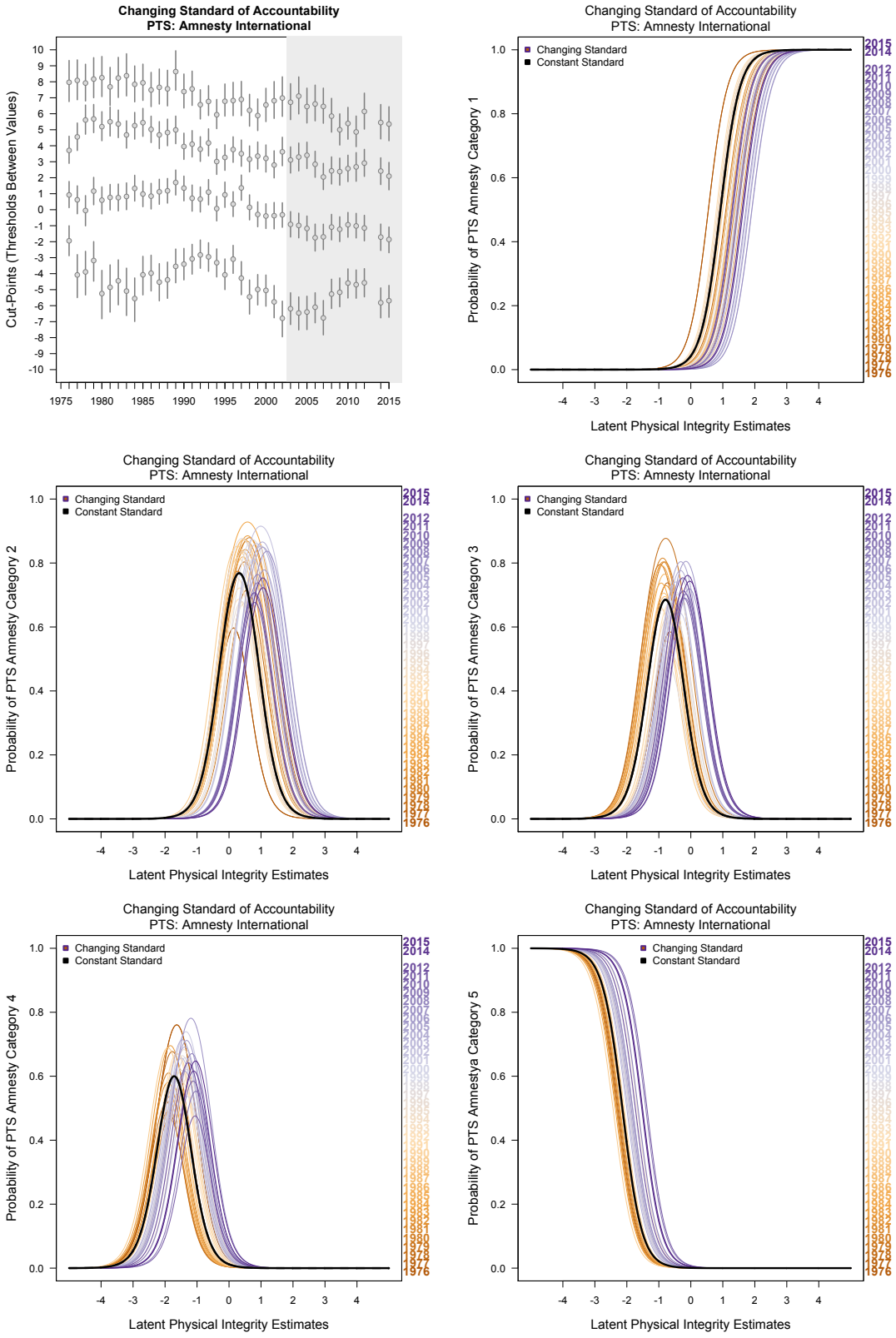


Figure 16: A decrease in the difficulty cut-points in the upper left panel translates directly into a change in the probability of being classified as a 1, 2, 3, 4, or 5 on the original PTS Amnesty Scale such that begin classified as 5 (e.g., frequent abuse) becomes more likely and 1 (e.g. no abuse) becomes less likely as a function of time.

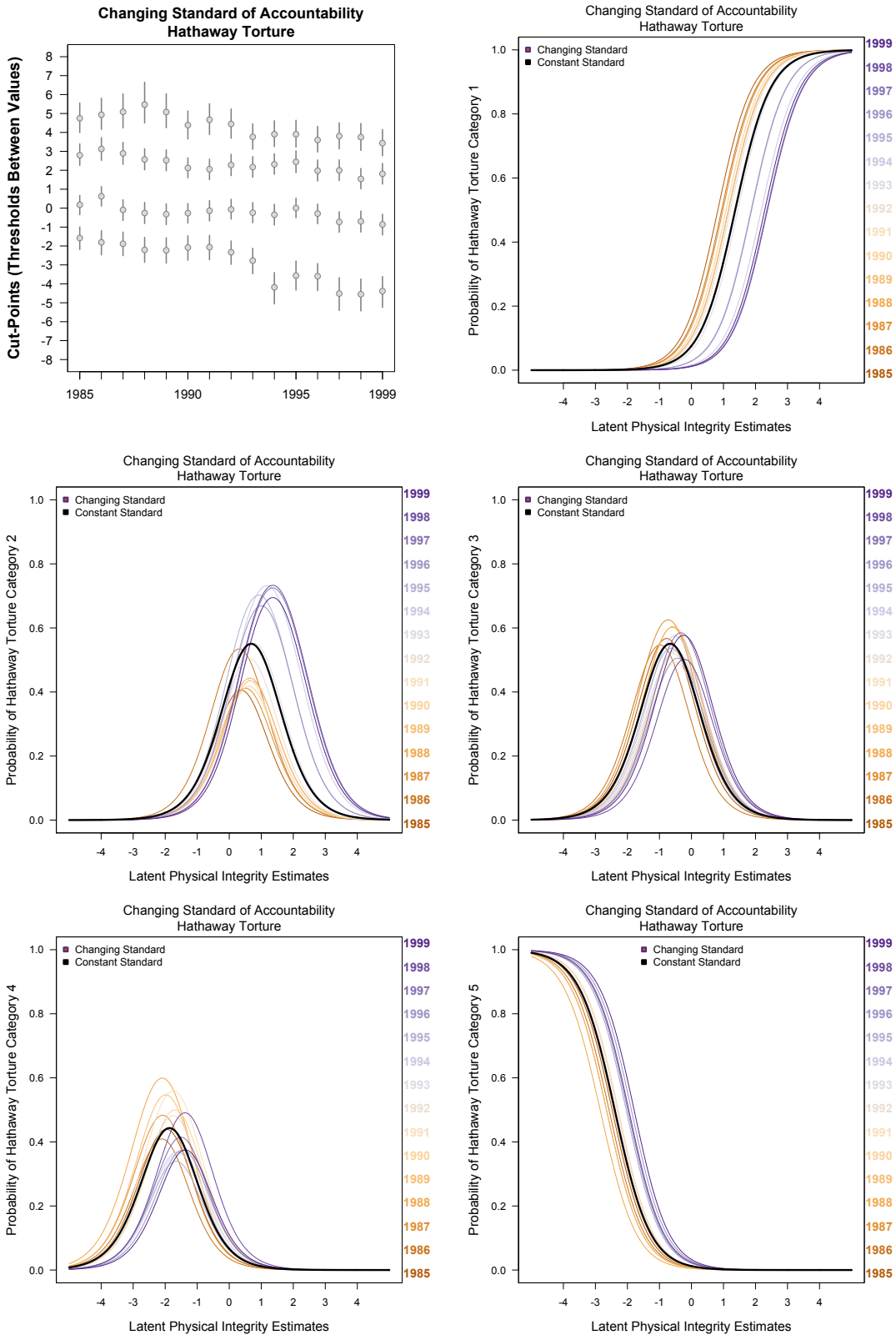


Figure 17: A decrease in the difficulty cut-points in the upper left panel translates directly into a change in the probability of being classified as a 1, 2, 3, 4, or 5 on the original Hathaway Scale such that begin classified as 5 (e.g., frequent abuse) becomes more likely and 1 (e.g. no abuse) becomes less likely as a function of time.

In summary, the item difficulty parameters determine the placement of the units along the latent variable range. The item discrimination parameters determine the level of precision in that placement. This type of probability based model is different from the construction of an additive scale. For an additive scale the item-value (the categorize 0, 1, or additional value), is also the value of the item-weight, which translates directly into a unit of the additive scale. It is therefore possible to discuss an additive scale in terms of the size of the item weights deterministically. That is, knowing the value of an item lets us know precisely the contribution or weight of that item to the additive scale. This is not how the latent variable model works. Instead, it builds probability models that relate the value of the latent trait for each unit, to the observed value of the item.

With an additive scale, it is possible to discuss the weight of the item using the same unit-of-measure as the value of the particular items. Getting a correct answer on a knowledge assessment, means that the student receives one additional point on the additive scale of their test score. Across all students, correctly answering the same question means getting one additional unit on the additive scale. It is not possible to directly translate values of the latent trait into values of the unit. However, we can consider, on average, the difference in means for country-year units, that are coded at one value on a human rights variable relative to the country-year units coded a different value. Since many of the human rights variables are ordered scales, I look at the differences in means for the latent trait for the units with the BEST score compared to the WORST score. Figure 18 graphs this differences in means graphically for three models. Because the models are nested with respect to units and items, it is possible to directly compare the difference in means across the three models. What the graphs reveal, is that on average the three latent variable models are assigning practically identical weight, in terms of the latent trait itself, to each of the 16 items (there are a few small differences that are probabilistically distinct from zero because the posterior values for these means are estimated with high precision).

In the constant standard model, the item difficulty and item discrimination parameters are estimated for all units. So there is only one inflection point for the binary variables, and a set of inflection points for the ordered variables. In the all varying standard model, the item difficulty are estimated for all units but the item difficulty parameters are estimated each year, so, though the precision of the inflection point is fixed over time, the location varies for all units. This makes it so that comparisons across time are

not possible. However, the average difference in means are going to be similar for this model compared to the others because, on average, all of the models use the same standard normal prior distribution on the latent trait to govern the placement of units. The all varying model resets this prior each year, but on average it is the same as the prior for constant standard model and the changing standard model.[2] As discussed in the main manuscript, over time comparisons are not possible. However, by averaging over all country-year units, it is possible to compare these difference in means which situate the average weight of the latent variable assigned to the average BEST and average WORST cases for each of the 16 observed human rights variables.

What these comparisons demonstrate is that the models are, on average, providing the same weight to each categorical grouping of country-year units. It is important to emphasize though, that the models are different because they do not treat each period of time the same way. We can see that using the same difference in means statistic. Instead of looking at the difference of means between the BEST and WORST value for one item, we can look at the difference in means for one time period compared to another. The difference in means for the latent variable for the period prior to 1990 compared to the period from 1990 forward for the changing standard of accountability model is 0.46 $[0.45,\ 0.47]$. For the constant standard model it is 0.06 $[0.05,\ 0.07]$ and for the all varying standard model it is 0.05$[\,0.04, 0.06]$. Thus, the item-weights, in terms of values of the latent variable estimates, are on average the same across the models but vary over time, depending on which model is used. The all varying standard model proposed by Cingranelli and Filippov (2018*a*) is not identified with respect to time, but again, the latent variable estimates are, by coincidence, very similar to the estimates from the constant standard model because it does not change over time either. The small difference for these two models is because of new units entering the model in later years. These new units tend to receive better scores on the observed human rights variables.

Overall, what these difference in means tests reveals is that one average, the models treat each of the variables the same. Differences between the estimates arise for some units relative to others only when the models account for time in a substantively meaningful way.

[2]There are some small differences in the variance between these three models but these between model comparisons are still possible by normalizing the variance across the three. This is not necessary for making comparisons within models or using the models to make predictions of other variables.

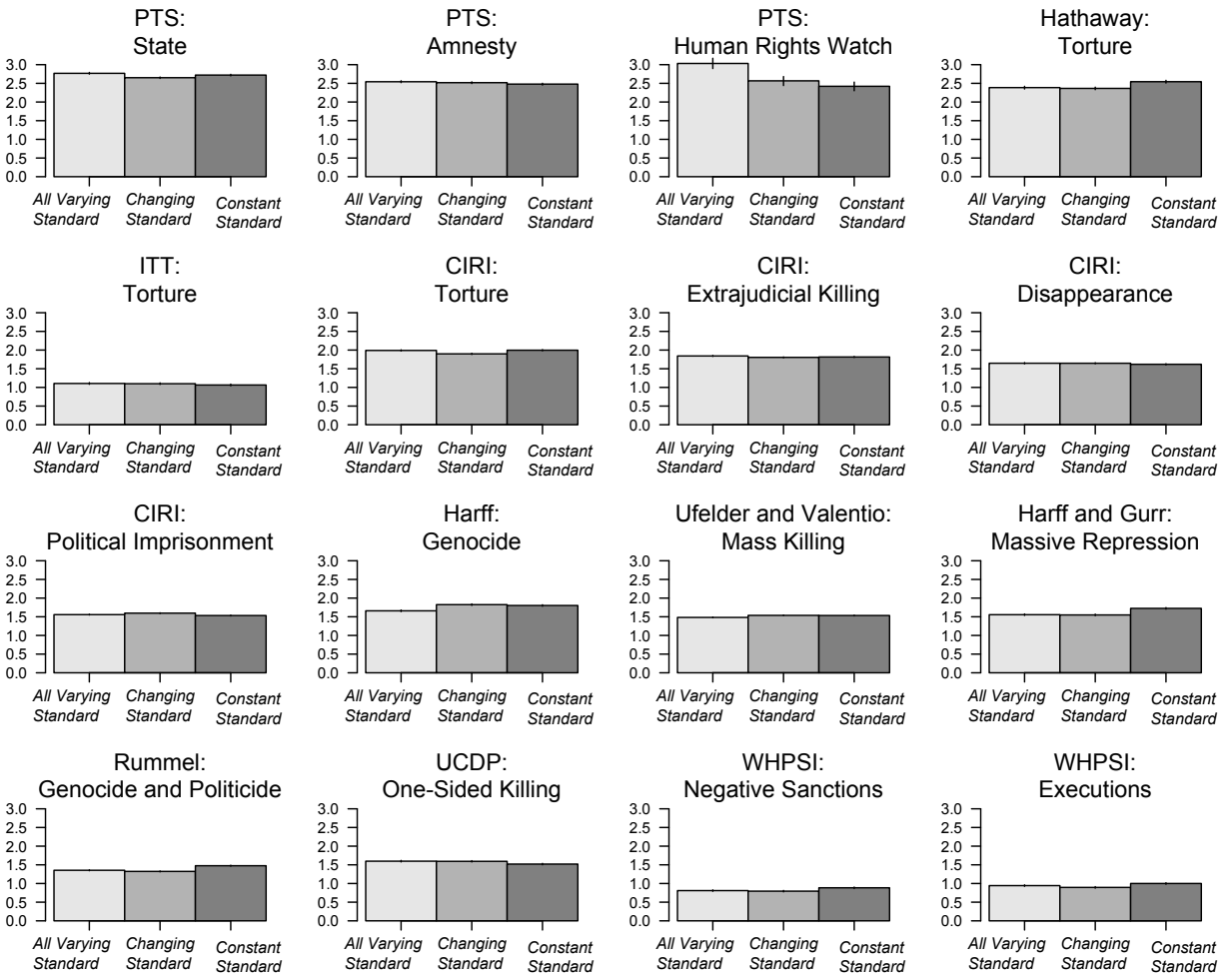


Figure 18: By averaging over all country-year units, it is possible to compare these difference in means which situate the average weight of the latent variable assigned to the average BEST and average WORST cases for each of the 16 observed human rights variables. What these difference in means tests reveals is that one average, the models treat each of the variables the same. Differences between the estimates arise for some units relative to others only when the models account for time in a substantively meaningful way.

# E Face-Validity and Concurrent-validity

Both face validity and concurrent validity are types of construct validity that relate some aspect of a measure, either its operational definition or its empirical content, to a theoretical concept. 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?[3] It is a specific check to assess whether there will be conceptual (translational) error in the resulting data generated by the operational protocol. 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*b*) 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 in mind. 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*)).

[3] 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).

# F Latent Variable Estimation and Extrapolation

The latent variable model is not an extrapolation. An extrapolation is a prediction of a value for a 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 the distribution of this unit specific uncertainty by year in Figure 19.

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*b*; Schnakenberg and Fariss, 2014). It is also consistent with advise from Bolck, Croon and Hagenaars (2004) and Mislevy (1991), which I discuss in Appendix section J below.

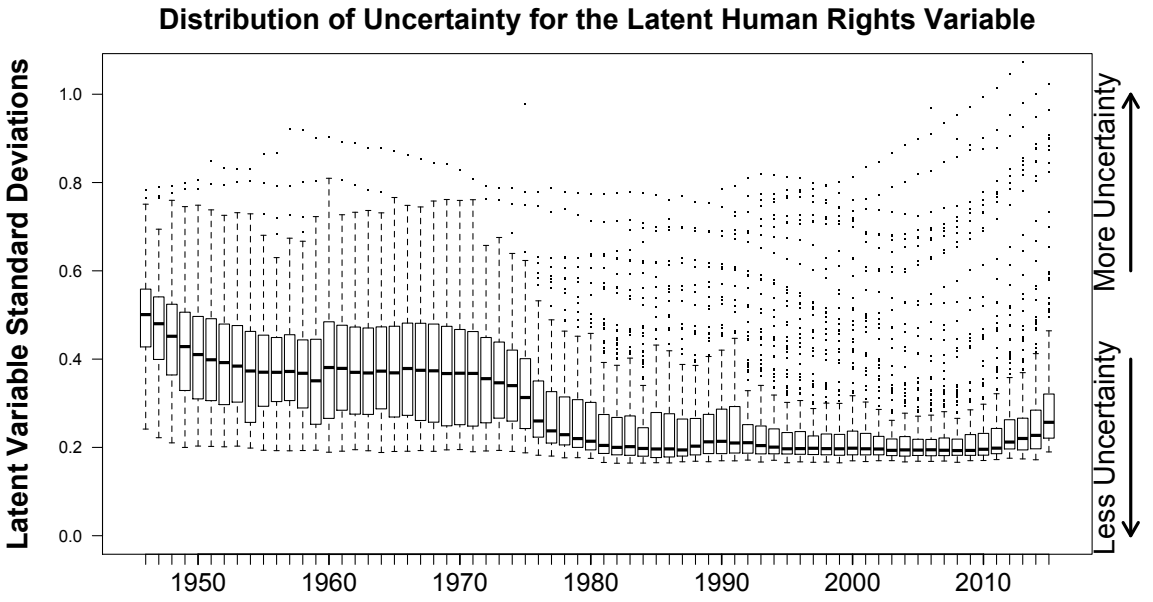


Figure 19: Modified from Fariss (2018*a*): The yearly distribution of the standard deviations from the latent variable estimates from 1946-2015. Though not every one of the repression variables is measured for each country-year unit, the latent variable model is able estimate a value of the latent variable 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*b*) for additional details.

# G The Comparative Method: Why Myanmar and Denmark are Comparable

Cingranelli and Filippov (2018*a*) ask "Why should mass killing in Myanmar, for example, have any effect on the calculation of the scores of countries such as Denmark?" 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 standard 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, based on categorical information that meets these criteria.

# H Deviant Case Studies

Cingranelli and Filippov (2018*a*) state that "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 the same authors (Cingranelli and Filippov, 2018*b*). In particular, Fariss (2018*a*) discusses the deviant case of the United States in 1953 (more on this case below). 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."

One example of a potentially deviant case, as Cingranelli and Filippov (2018*b*,*a*) suggest, is the United States in 1953. First we should consider why it is placed on the low end of the latent variable. For two of the event-based variables, the US was coded as repressive for specific reasons: the US engaged in political killings during the 1950s and 1960s in the American South and it executed two Soviet spies in 1953. These are not trivial matters. These events do not even pick up the investigations into Communist activists by Senator Joseph McCarthy that were also taking place in the early 1950s. Of course, monitors and the media may be more aware of these events because of the high levels of press freedom in the US

relative to other countries. This is an example of the challenges in modeling human rights respect, and it is these challenges that the latent variable model helps to address.

To further address this issue, 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.[4] 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.

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

# I Latent Variable Trends for Democracy and Non-Democracy over Time

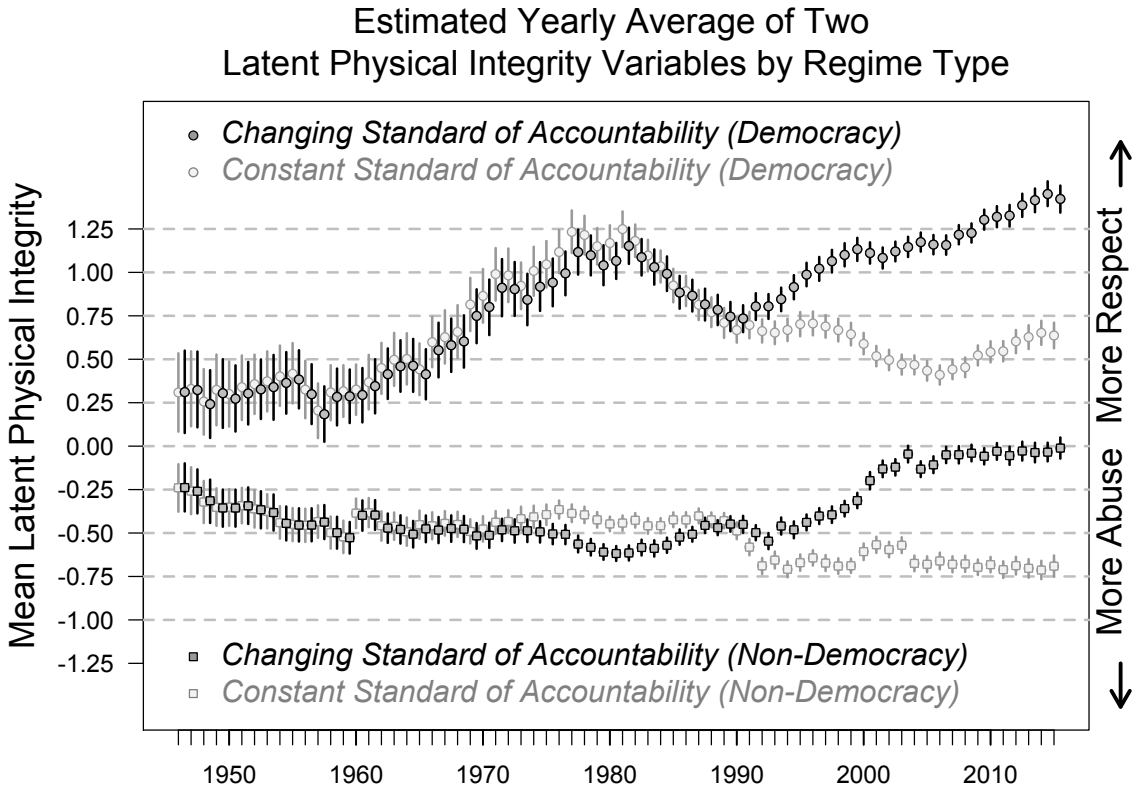


Figure 20: Modified from Fariss (2018*a*): The graph displays yearly mean and credible intervals for these same variables across democratic and non-democratic states as measured by Polity IV (values of 6 or greater). Only the latent variable estimates that assume a changing standard of accountability 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*b*) for additional details.

# J Using the Latent Variable Estimates in Applied Research

Cingranelli and Filippov (2018*a*) conclude that "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)" (pg. 7).

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 is a scale that is also 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*b*) also discusses this in Appendix section M.

Fariss (2014*b*) also provides several additional methodological suggestions for using the latent variable estimates or the original categorical variables from PTS or CIRI in the conclusion of that article:

> "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" (314).

# 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.

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

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.

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

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

Cingranelli, David L. and Mikhail Filippov. 2018*b*. "Problems of Model Specification and Improper Data Extrapolation." *British Journal of Political Science* 48(1).

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

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

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

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

Fariss, Christopher J. 2014*a*. Replication Data for: Exploring the Dynamics of Latent Variable Models. Technical report.
**URL:** *https://dx.doi.org/10.7910/DVN/25830*

Fariss, Christopher J. 2014*b*. "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.

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

Gibney, Mark and Matthew Dalton. 1996. The Political Terror Scale. In *Human Rights and Developing Countries*, ed. D. L. Cingranelli. Vol. 4 of *Policy Studies and Developing Nations* Greenwich, CT: JAI Press pp. 73–84.

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.

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

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.

Marshall, Monty G., Ted R. Gurr and Barbara Harff. 2009. "PITF - STATE FAILURE PROBLEM SET: Internal Wars and Failures of Governance, 1955-2009." *Dataset and Coding Guidelines* .

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

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

Rummel, Rudolph J. 1994. "Power, Geocide and Mass Murder." *Journal of Peace Research* 31(1):1–10.

Rummel, Rudolph J. 1995. "Democracy, power, genocide, and mass murder." *Journal of Conflict Resolution* 39(1):3–26.

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.

Sundberg, Ralph. 2009. Revisiting One-sided Violence: A Global and Regional Analysis. In *States in Armed Conflict*, ed. Lotta Harbom and Ralph Sundberg. Uppsala: Universitetstryckeriet.

Taylor, Charles Lewis and David A. Jodice. 1983. *World Handbook of Political and Social Indicators Third Edition*. Vol. 2, Political Protest and Government Change. New Haven: Yale University Press.

Trochim, William M.K. and James P. Donnelly. 2008. *Research Methods Knowledge Base*. 3rd ed. Mason, OH: Atomic Dog.

Ulfelder, Jay and Benjamin Valentino. 2008. "Assessing Risks of State-Sponsored Mass Killing." http://dx.doi.org/10.2139/ssrn.1703426.

Wayman, Frank W. and Atsushi Tago. 2010. "Explaining the onset of mass killing, 1949–87." *Journal of Peace Research* 47(1):3–13.