American Political Science Review (2018) 112, 4, 1083–1089
doi:10.1017/S0003055418000254 © American Political Science Association 2018

Letter

Are Human Rights Practices Improving?

DAVID CINGRANELLI *Binghamton University*
MIKHAIL FILIPPOV *Binghamton University*

Has government protection of human rights improved? The answer to this and many other research questions is strongly affected by the assumptions we make and the modeling strategy we choose as the basis for creating human rights country scores. Fariss (2014) introduced a statistical model that produced latent scores showing an improving trend in human rights. Consistent with his stringent assumptions, his statistical model heavily weighted rare incidents of mass killings such as genocide, while discounting indicators of lesser and more common violations such as torture and political imprisonment. We replicated his analysis, replacing the actual values of all *indicators of lesser human rights violations with randomly generated data, and obtained an identical improving trend. However, when we replicated the analysis, relaxing his assumptions by allowing* all *indicators to potentially have a similar effect on the latent scores, we find no human rights improvement.*

Science is advanced by a community of investigators who often disagree about explanations for important phenomena. Sometimes disagreements over fundamental conceptual and theoretical issues are so deep that researchers reach an impasse. In statistical analysis, they may disagree over what evidence should be used or how various types of indicators should be weighted to measure important concepts. In that case, depending upon the indicators each scientist emphasizes, conflicting findings persist.

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) suggested a novel statistical approach to reevaluate human rights indicators. The new measure he introduced gave rise to further debate. Most importantly, 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).

Fariss's new scores encourage scholars to reexamine many research findings accumulated by the subfield. Since his scores, on average, increase over time, they are likely to be correlated with many variables that also increase over time such as treaty ratifications, degree of globalization, the degree of economic inequality within nations, and democratization. Fariss (2014, 2018) has already challenged previous results on the effects of human rights treaty ratification (but see Cingranelli and Filippov 2018). Future studies using his scores are likely to produce many findings that conflict with previous results.

Fariss's model specification and results were strongly affected by his assumptions. He assumes that:

- Mass killing events (such as genocide) and lesser human rights violations are indicators of the same underlying variable—respect for physical integrity human rights.
- Incidents of mass killing are recorded more accurately than lesser violations.
- 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" (S_1 varies) in human rights reports of lesser human rights violations.
- There has been no change in the "standards of accountability" (S_2 = constant) in records of mass killings.
- 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 ($S_1 - S_2$) between the two types of records.

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 (S_2 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.

Fariss's assumptions, and, crucially, that S_2 does not vary, led him to use a statistical technique that heavily weights rare "event-based" incidents of mass

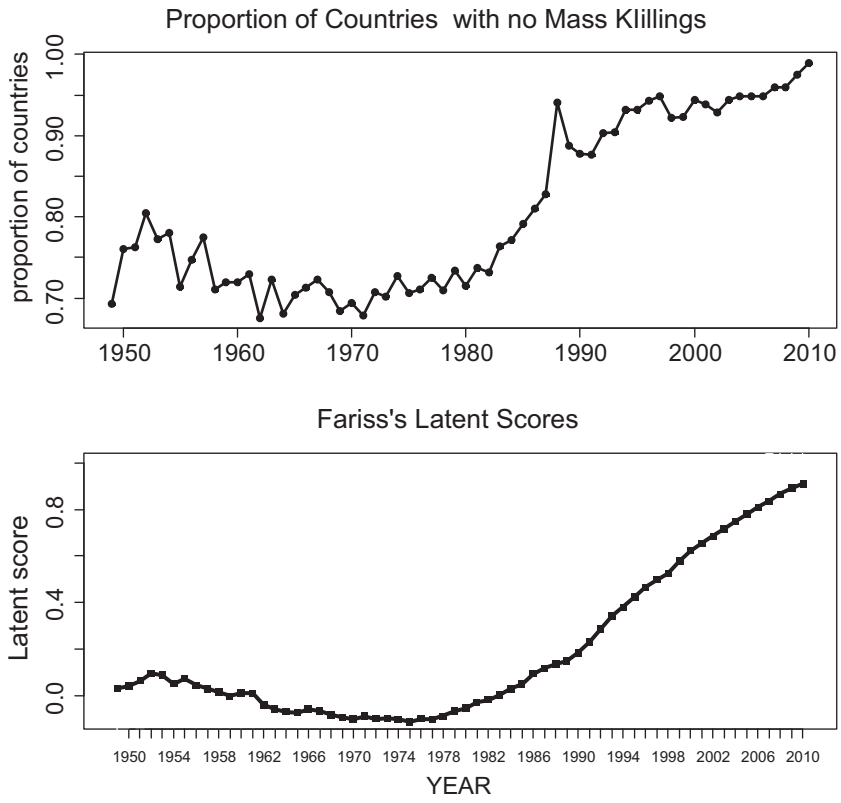
David Cingranelli is a Professor of Political Science, Binghamton University, State University of New York, Vestal Parkway East, Binghamton, NY 13902-6000, USA (davidc@binghamton.edu).
Mikhail Filippov is an Associate Professor of Political Science, Binghamton University, State University of New York, Vestal Parkway East, Binghamton, NY 13902-6000, USA (Mikhail.filippov@gmail.com).

The authors thank Rodwan Abouharb, Sabine Carey, David Davis, Peter Haschke, Neil Mitchell, and David Richards for their helpful comments on earlier versions of this paper. Replication files are available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/KGVBNC.

Received: February 2, 2017; revised: December 17, 2017; accepted: April 27, 2018. First published online: June 13, 2018.

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

FIGURE 1. A Comparison of the Trends in Mass Killing Events and in Fariss's Latent Scores, 1949–2010



killing such as genocide, discounting "standards-based indicators" of "lesser" and more common human rights violations such as torture and political imprisonment. 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).

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.

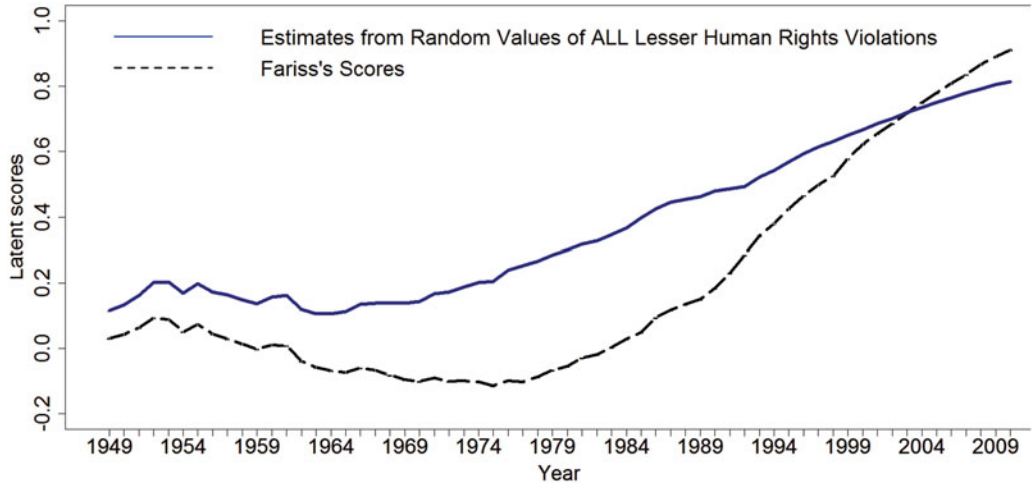
More generally, we show that alternative modeling choices have substantive consequences for answering questions about human rights improvement and for the development of human rights theory and relevant public policy. Any modeling strategy must assign weights to different types of evidence. We do not 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.

Before discussing the substantive consequences of modeling decisions as they relate to human rights, it is illustrative to consider one of the most famous similar divides in political science. This was the debate over whether the power structure within US communities was pyramidal (elitist) or horizontal (pluralist). Scholars who used the decisional method pioneered by Dahl (1961) discovered a power structure that was pluralist. Dahl and his associates attended public meetings in a particular city, recording who was influential and who was not. Others used the reputational method pioneered by Hunter (1953). They asked people in a community to identify the most influential people who, either formally or informally, swayed the outcomes of decisions. They found that community power structure was elitist. By 1980, research on this topic came to a halt with no satisfactory resolution of the disagreements. Neither conclusion was right or wrong. Each was based on a different theoretical position and modeling strategy. The choice of modeling strategy determined the conclusions.

The human rights subfield is engaged in a similar 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

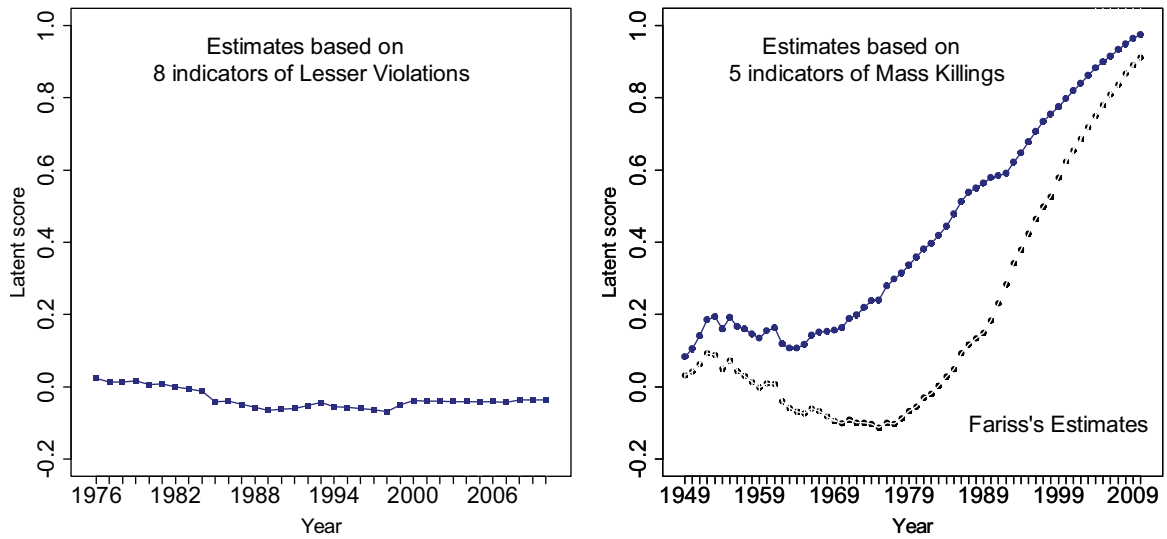
https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

FIGURE 2. A Comparison of the Trends of the Dynamic Latent Human Rights Estimates, 1949–2010



Black Line: Replication of Figure 3 in Fariss (2014, 308).
Blue Line: The values of ALL indicators of Lesser Human Rights violations are replaced by *random numbers.*

FIGURE 3. A Comparison of Latent Human Rights Trends Estimated Using Only Indicators of Lesser Human Rights Violations or Only Mass Killing Indicators



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.

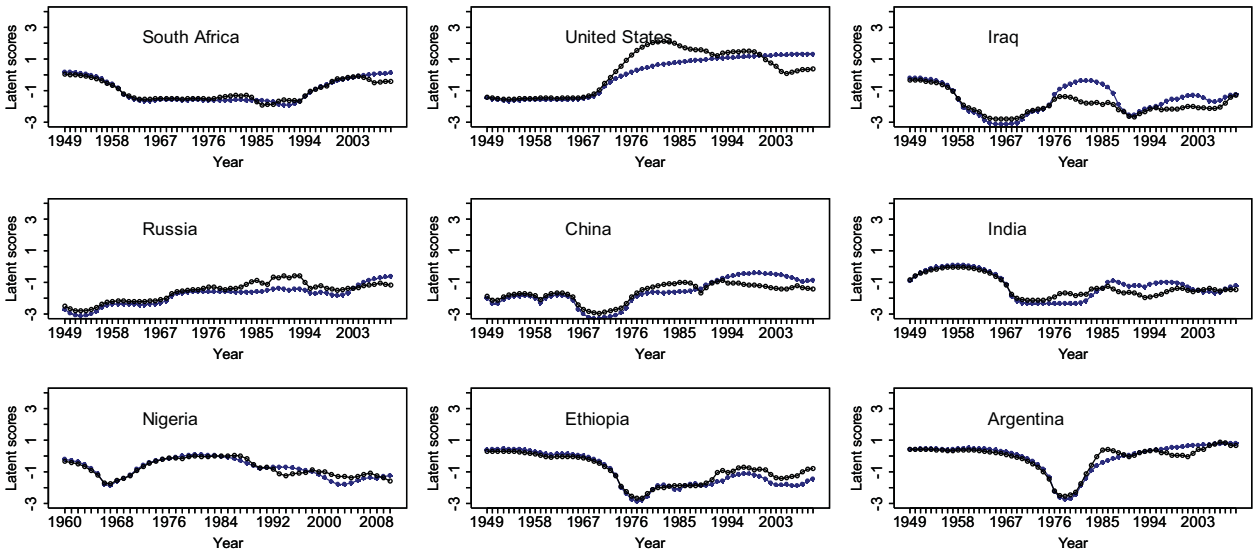
A NEW VERSION OF DYNAMIC IRT

The two most commonly used measures of human rights are the Political Terror Scale (PTS) and the CIRI Physical Integrity Index. Both data projects assign numerical scores to countries based on information included in annual reports produced by the US Department of State and Amnesty International. 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 led coders to make more negative assessments of human rights practices; (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). For counterarguments and contrary evidence, see Richards (2016) and Haschke and Gibney (2017).

Fariss suggested that possible biases in human rights data could be identified and corrected by estimating a

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

FIGURE 4. Fariss's Scores (unfilled circles) Compared to Latent Scores Estimated Using Only Five Indicators of Mass Killing, 1949–2010



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.

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 killing 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).[1] 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) 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). 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.

[1] We generated the random numbers by several distinct algorithms. All methods produced similar improving trends. See online appendix.

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. For more country examples, see our online appendix (Figures A.1.1– A.1.4).

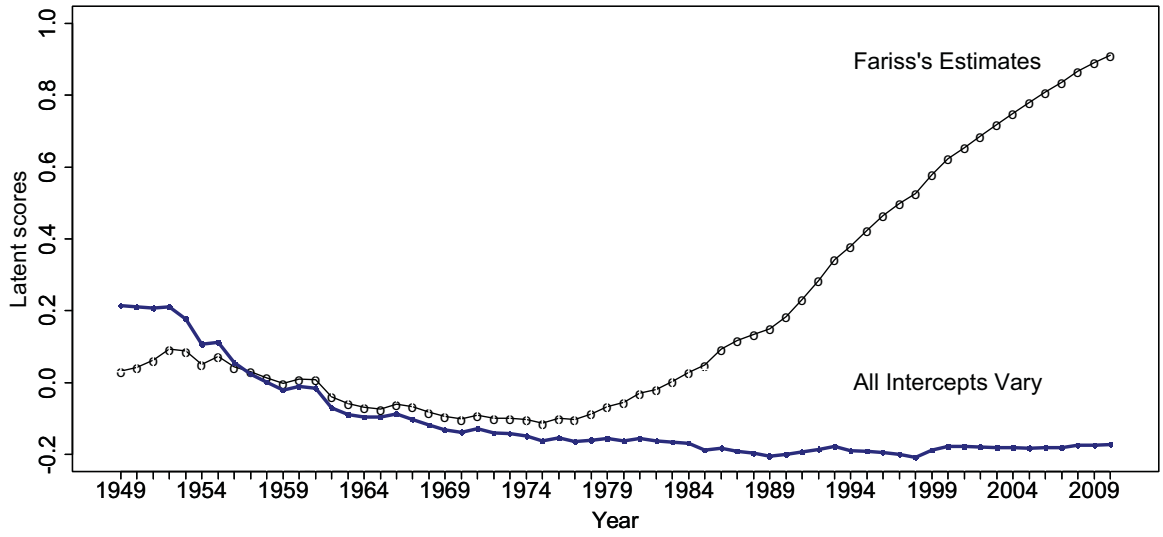
MODELING CHOICE MATTERS

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, Berger, and Burdick 2013). 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.

When Fariss's model combines the two types of data in a single estimator, it treats the two groups of indicators differently (Fariss 2014, 305–306). It sets the mass killing indicators to follow a logistic regression with a fixed intercept (cut point) but allows the indicators of lesser human rights violations to follow an ordered logistic regression with variable intercepts for every year. Thus, the latent variable has 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 is much easier for the algorithm to fit the latent variable as there are several dozen additional parameters (time specific intercepts) that could also adjust. Consequently, variation in the mass killing indicators generates the improving trend of the latent variable. This is

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

FIGURE 5. Means of Dynamic Latent Physical Integrity Estimates, 1949–2010



Key: Unfilled Circles: Replication of Figure 3 in Fariss (2014, 308).
Filled Circles: The scores are calculated with all indicators assumed to have a similar relationship to the latent variable; that is, all intercepts are allowed to vary. The assumption is that there also has been a change in the standards for recording mass killings.

why the actual values of the indicators of lesser human rights violations do not matter much.

Fixing the intercepts for mass killing indicators is necessary to obtain the scores showing an improvement in human rights. A model where the intercepts for all items vary across time produces latent variable estimates similar to those from a model where none of the intercepts vary. When we rerun 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.

MASS KILLINGS AS A BASELINE?

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 difference in the changes in the standards of accountability ($S_1 - S_2$) 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.

Though Fariss's model distinguishes between events-based and standards-based data, it is important to recognize that even mass killing indicators are standards-based. 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.

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.

Even if the standards for recording mass killings had been more consistent over time and among coders,

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

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.

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?

As a consequence of putting so much weight on rare mass killings, for many countries, the trends in Fariss's scores are nonintuitive, 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.

Particularly questionable are the trends in the scores between 1949 and 1975. No records of lesser human rights violations are available for this period. Moreover, for the majority of country-years between 1949 and 1976, the actual codes for all 13 human rights indicators before extrapolation were zero. Yet all country-years were assigned different latent human rights scores.[2]

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. Fariss's scores place 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 (Figures A.2.1– A.2.5).

[2] Their credible intervals might overlap, indicating no statistically significant difference.

CONCLUSION

Has government protection of physical integrity rights improved? The answer depends on the evidence one chooses to emphasize. Fariss's model, which accentuates the decline in mass killing events, produces scores and trends suggesting that the answer is *yes*. We and others, who have focused on lesser violations of physical integrity such as torture or political imprisonment, acknowledge that a smaller proportion of countries are experiencing mass killings, but still think the answer is *no* or at least *we do not know.*

Fariss (2014) has challenged research in the field by highlighting the contrasting conclusions we would reach about human rights improvement if we considered the striking decline in records of mass killings. His model 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. Computationally, the trend in Fariss's scores mirrors the trend in mass killing events. 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 as shown in Figure 5 above.

Fariss's model 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 the intercepts for others were not.

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 could not suffer from changing standards of accountability. Yet, our simulations show that Fariss's model specification would "correct" random values, too.

We conclude with some practical advice for scholars and policy analysts in the human rights subfield. Those who use Fariss's scores 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.

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press

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).

Finally, we agree with Fariss (2014, 303) 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." 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.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055418000254

Replication materials can be found on Dataverse at: https://doi.org/10.7910/DVN/KGVBNC.

REFERENCES

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

Cingranelli, David, and Mikhail Filippov. 2018. "Problems of Model Specification and Improper Data Extrapolation." *British Journal of Political Science* 48 (1): 273–74.

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

Clark, Ann Marie, and Kathryn Sikkink. 2013. "Information Effects and Human Rights Data." *Human Rights Quarterly* 35 (3): 539–68.

Dahl, Robert. 1961. *Who Governs*. New Haven, CT: Yale University Press.

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

Fariss, Christopher. 2018. "The Changing Standard of Accountability and the Positive Relationship Between Human Rights Treaty Ratification and Compliance." *British Journal of Political Science* 48 (1): 239–71.

Gutiérrez-Sanín, Francisco, and Elizabeth Wood. 2017. "What Should We Mean by 'Pattern of Political Violence'?" *Perspectives on Politics* 15 (1): 20–41.

Hafner-Burton, Emily, and Jaames Ron. 2009. "Seeing Double: Human Rights Impact through Qualitative and Quantitative Eyes." *World Politics* 61 (2): 360–401.

Harff, Barbara, and Ted Robert Gurr. 1988. "Toward Empirical Theory of Genocides and Politicides." *International Studies Quarterly* 32 (3): 59–71.

Haschke, Peter, and Mark Gibney. 2017. "Are Global Human Rights Conditions Static or Improving." In *Peace and Conflict*, eds. David Backer, Ravi Bhavnani, and Paul Huth. New York: Routledge Press, 87–97.

Hill, Daniel. 2016. "Democracy and the Concept of Personal Integrity Rights." *Journal of Politics* 78 (3): 822–35.

Hunter, Floyd. 1953. *Community Power Structure*. Chapel Hill: University of North Carolina Press.

Martin, Andrew, and Kevin Quinn. 2002. "Dynamic Ideal Point Estimation Via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10 (2): 134–53.

Poe, Steven, and Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s." *American Political Science Review* 88 (4): 853–72.

Richards, David. 2016. "The Myth of Information Effects in Human Rights Data." *Human Rights Quarterly* 38 (1): 477–92.

Rummel, Rudolph. 1994. *Death by Government*. New Brunswick, NJ: Transaction Press.

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

Wang, Xiaojing, James Berger, and Donald Burdick. 2013. "Bayesian Analysis of Dynamic Item Response Models in Educational Testing." *The Annals of Applied Statistics* 7 (1): 126–53.

Wood, Reed, and Mark Gibney. 2010. "The Political Terror Scale (PTS): A Re-Introduction and a Comparison to CIRI." *Human Rights Quarterly* 32 (2): 367–400.

https://doi.org/10.1017/S0003055418000254 Published online by Cambridge University Press