RESEARCH

# the construction of human rights: accounting for systematic bias in common human rights measures

*mark david nieman*[a, *] *and jonathan j. ring*[b]

[a]Department of Political Science, Iowa State University, 537 Ross Hall, Ames, Iowa 50011, USA.

E-mail: mdnieman@iastate.edu

[b]Department of Political Science, University of Michigan, 6642 Haven Hall, Ann Arbor, Michigan 48109, USA.

E-mail jonring@umich.edu

*Corresponding author.

doi:10.1057/eps.2015.60

**Abstract**

Empirical human rights researchers frequently rely on indexes of physical integrity rights created by the Cingranelli-Richards (CIRI) or the Political Terror Scale (PTS) data projects. Any systematic bias contained within a component used to create CIRI and PTS carries over to the final index. We investigate potential bias in these indexes by comparing differences between PTS scores constructed from different sources, the United States State Department (SD) and Amnesty International (AI). In order to establish best practices, we offer two solutions for addressing bias. First, we recommend excluding data before 1980. The data prior to 1980 are truncated because the SD only created reports for current and potential foreign aid recipients. Including these data with the more systematically included post-1980 data is a key source of bias. Our second solution employs a two-stage instrumented variable technique to estimate and then correct for SD bias. We demonstrate how following these best practices can affect results and inferences drawn from quantitative work by replicating a study of interstate conflict and repression.

**Keywords** bias; human rights; truncation; instrumental variables

The empirical human rights literature frequently relies on indexes of physical/personal integrity rights from one of two major data sets: Cingranelli-Richards (CIRI) or the Political Terror Scale (PTS) (Cingranelli and Richards,

2010; Gibney and Dalton, 1996). Both of these projects draw from United States (US) State Department (SD) reports and Amnesty International (AI) reports. The data are frequently used without questioning the validity of the measures. It is important to keep in mind, however, that the State Department is an agency that helps formulate the foreign policy of the most powerful state in the international system, while Amnesty International is a globally connected non-governmental organisation (NGO) headquartered in London, which boasts 3 million supporters with activity in 150 countries. By publishing their human rights reports, both organisations are actively seeking to shape the preferences of individuals, states, international organisations, and NGOs throughout world.

Despite each relying on expert coded indexes drawn from SD and AI reports, there are several important differences between CIRI and PTS. For one, CIRI is an additive composite of different human rights components, while PTS is a holistic report of the overall condition of human rights. CIRI uses both SD and AI reports to produce one scale, while PTS produces two separate scales from the SD and AI reports. An unanswered question regarding PTS is how the two scales should be used. One strategy is to simply average the two scores. Although this is alluded to in Wood and Gibney (2010), it is not specifically endorsed by the PTS directors. Yet, the average is included pre-calculated from the PTS website. Any systematic difference, or bias, contained within a component used to create the CIRI or PTS measures such as SD scores, also carries over into the final index. We find evidence of systematic differences between SD and AI reports when the data are disaggregated by time period and geographical region. More generally, we find that these differences reflect US geo-political concerns. Thus, we treat bias as the extent to which SD scores deviate from AI scores.[1]

In order to establish best practices for quantitative human rights researchers, we offer two solutions for addressing this systematic bias. First, because data from early SD reports are truncated – the SD only created reports for current and potential foreign aid recipients until 1980 – there is clear selection bias present within this subset of data. Rather than including this data with the broader, more systematically collected post-1980 human rights data, we recommend either excluding the pre-1980 data or imputing these data using our second solution. Our second solution is to employ a two-stage instrumented variable technique to estimate and then correct for SD bias. We demonstrate how following these best practices can affect results and inferences drawn from quantitative work by replicating a study of interstate conflict and repression.

## DIFFERENCES IN STATE AND AMNESTY PTS SCORES

The statistical study of human rights has become a burgeoning field over recent decades. The widespread use of measures that account for a state's repressive behaviour towards its citizens has created a core set of empirically supported theoretical assertions. These include that democracies are less repressive than non-democracies, war increases repression, and that common law legal systems are associated with less repression (e.g., Hill and Jones, 2014). While there may be new discoveries concerning new covariates that explain why human rights are violated, we believe a great deal of utility can now be gained by focusing on the accuracy of the estimates.[2] Should state A enact policy X or Y to help improve the human rights in state B? Asking this question moves beyond identifying the causes of human rights violations to the relative importance of the various causes. In other words, through the calculation

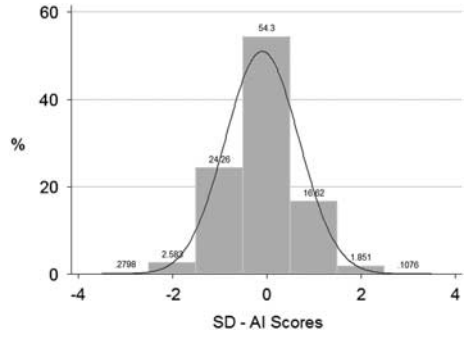


**Figure 1** Distribution of $PTS_{SD}$–$PTS_{AI}$.

and interpretation of substantive effects, we can compare the effect of $X_1$ with $X_2$. Asking more nuanced questions about the relative importance of rival causes requires greater faith in our measures. Given this shift in scholarly focus, we believe it is time to provide a more critical evaluation of the measures the discipline uses to make these inferences.

In looking at these issues we are concerned with establishing best practices for using any human rights indexes using SD reports. In this study, we focus on PTS, one of the most widely used measures of personal integrity.[3] PTS is a five-point index describing the human rights conditions in a particular country in a given year.[4] For each country-year, expert coders on the PTS project assign a single score based on their reading of the State Department reports ($PTS_{SD}$) and an independent score based on Amnesty International reports ($PTS_{AI}$). The scores range from one to five with higher scores indicating more repression. For a careful review of the construction of the scores, see Wood and Gibney (2010).

There have been two previous attempts to systematically examine differences in PTS human rights scores from the State Department and Amnesty International reports: Poe *et al* (2001) and Qian and Yanagizawa (2009). Each article starts from a common criticism that the SD provides biased reports. Poe *et al* (2001: 651) suggest the potential that the SD unfairly paint[s] with the tar of repression countries ideologically opposed to the US, while unjustly favouring countries where the US has had a compelling interest'. To investigate this claim, both Poe *et al* (2001) and Qian and Yanagizawa (2009) subject the difference between the PTS scores from SD and AI reports to regression analyses.

The difference variable, $PTS_{SD}$–$PTS_{AI}$, has a theoretical range of −4 to 4.[5] Higher numbers indicate that PTS coders have assigned a higher score to a country-year based on their reading of the SD reports relative to the Amnesty International report.[6] As long as PTS coders are consistent in the application of the coding rules, this means that the differences in the portrayal of human rights conditions are coming from the reports themselves. Thus, the difference variable is capturing the bias of the SD relative to Amnesty International.

Poe *et al* (2001) analyse 20 years of data over the period 1976–1995. They initially present descriptive statistics, and we replicate that exercise while adding PTS data from 1996–2013. Our initial sample contains 4,646 country-years compared with 2,331 in the original study, nearly doubling the time-series-cross-sectional space. Figure 1 presents the distribution of the difference variable from our sample. Zero is the modal category, which means that most of the time, PTS coders assign the same score to a country regardless of whether they are looking at SD reports or AI reports. We find what looks generally like a normal distribution, much more evenly dispersed in the full sample than in the shorter sample.[7] Furthermore, around 95 per cent of the cases fall within an absolute value of one. This number is nearly identical to Poe *et al* (2001), but there is a small change. Of the 5 per cent that have a difference of two or more points, we find that SD is getting the lower (more favourable HR conditions) score about one and a half times more

frequently (2.86 per cent of the total sample compared with 1.96 per cent). In Poe *et al's* (2001) sample, when there is a difference of 2 or more, $PTS_{SD}$ is the lower score nearly four times as often as $PTS_{AI}$ (3.22 per cent compared with 0.82 per cent). This trend in the tails of the distribution is also evident closer to the centre, where most of the data are located. In Poe *et al's* (2001) sample, when looking at scores that differ by an absolute value of one, $PTS_{SD}$ was lower in 29 per cent of all cases compared with only 12.2 per cent of cases in which $PTS_{AI}$ was lower. In the full sample, this is much closer to even with 24.3 and 16.6 per cent respectively. Overall, we conclude that the bias found by Poe *et al* (2001) when focusing on 1976–1997 is still evident in the full sample (1976–2013), but it has significantly decreased with the inclusion of later years.

Poe *et al* (2001) and Qian and Yanagizawa (2009) each conduct multivariate regression analyses on the difference in PTS scores. A common finding between the two studies is that US allies receive more lenient human rights reports by the SD relative to AI, though Qian and Yanagizawa (2009) find this effect is reduced after the Cold War. Another common feature is that each employ unbalanced samples. The unbalanced nature of the data is not just because of new states entering the system at the end of the Cold War. Importantly, it is also because of systematically limited SD coverage before 1980. As (Poe *et al*, 2001: 654) note:

> *Publication of the reports began in 1976, as a means for Congress to keep tabs on recipients of US aid in an attempt to verify the wishes of Congress were being followed, but by 1980 the reports were covering a much more comprehensive set of UN member countries.*

Moreover, political closeness to the US is highly correlated with whether a state is included in SD reports before 1980. The reported findings are likely dampened owing to this selection bias in the pre-1980 data.

The results reported in the two articles cast doubt on research that uncritically employ PTS and CIRI scores of human rights. If SD reports are biased, and that bias is systematically related to covariates used to explain human rights behaviour, the analyses could report results over- or under-estimating their true effect. It may also affect the reported signs or significance levels, inferences drawn, or even whether a manuscript is seen as publishable (Esaray and Danneman, 2014; Neumann and Graeff, 2013).

## SOURCES OF BIAS

There are two types of bias potentially affecting PTS scores: systematic and random. Systematic bias results from strategic incentives on the part of the SD to provide more (less) favourable reports regarding a state's human rights practices. Random bias stems from the idiosyncratic nature of country offices and the SD's policies requiring personnel to change locations after 2 years (US Department of State, 2012). In contrast, AI has no such formal requirement. Such rapid personnel turnover introduces greater variation in human rights standards by SD coders, resulting in less reliable intra-country coding. For example, Figure 2 reports human rights scores for Togo from the SD in comparison with human rights scores from AI. It seems unlikely that the human rights practices in Togo really changed as dramatically as the reports issued by the SD indicated to PTS coders.

Random bias increases the error around parameter estimates in regression analyses using data generated by the SD. The added 'noise' from random bias reduces our confidence in parameter estimates. It does not, however, bias the

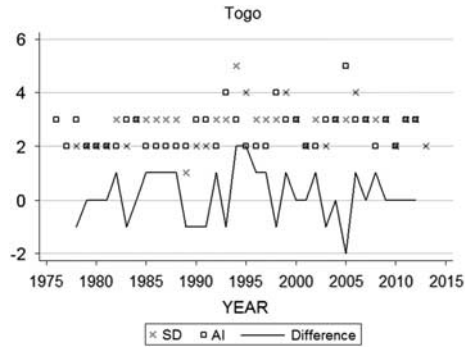


**Figure 2** Amnesty international and SD human rights scores for togo.

coefficients themselves. Systematic bias, on the other hand, does affect our coefficients and, thus, impacts the inferences that scholars make. For this reason, we focus on the sources of systematic bias.

## STRATEGIC INTERESTS AS A SOURCE OF SYSTEMATIC BIAS

What causes systematic bias in SD reports? Poe *et al* (2001: 657) suggest two explanations for systematic differences between $PTS_{AI}$ and $PTS_{SD}$ scores. First, the two organisations have different organisational missions. The US 'pursues power, and thus weighs security concerns more heavily than the human rights of non-Americans abroad' while the goals of AI are 'to forward the cause of human rights, worldwide'. Second, the US in general and the SD in particular have concerns other than human rights while AI is solely devoted to human rights. Since the US is concerned with maintaining the structure of the international system, it often supports norms of sovereignty at the expense of promoting human rights. Thus, 'it may have a tendency to treat more lightly than Amnesty International, so as not to interfere unduly in the affairs of other governments' (Poe *et al*, 2001: 657). Thus, the SD has incentives to look past alleged misdeeds by 'friends' and other states with high geo-political value in order to shelter them (and the US) from public scrutiny and criticism.

More generally, great powers often seek to expand their influence for either explicit strategic concerns or normative, ideological causes. The US and Soviet Union, for example, each emerged as superpowers with expansionary, messianic visions – the former promoting liberal, free-market capitalism and the latter centrally planned communism – and other states became increasing reliant on them for military and political guidance. Cuba, East Germany, and North Korea, for instance, heavily relied on Soviet aid, while West Germany, the Philippines, and South Korea entrusted their military security to the US (Westad, 2005).

In return for economic, military, and political aid, members of each bloc were obliged to undergo certain tasks in the name of their beneficiary. For instance, as a US ally, Turkey was expected to permit the placement of US troops and missiles within their borders, and to maintain relatively peaceful relations with Greece, a state considered by Turkey to be a bitter rival (Poe and Meernik, 1995; Meernik *et al*, 1998). Relative peace with Greece was important to the US in order to keep both within the Western bloc. Other aid recipients supported neighbouring rebel groups against their central governments, as in the case of the US-supported Contras in Nicaragua (based in Honduras) and the Soviet-supported FAR and MR-13 in Guatemala (assisted by Cuba), as part of the 'proxy wars' fought between the two superpowers (Kalyvas and Balcells, 2010; Westad, 1992). Moreover, aid recipients may also be expected to prevent potential ideological foes of their friendly superpower from taking power in their own country (Blanton, 2000; Poe and Meernik, 1995).

The degree of superpower influence produces observable effects on the behaviour of minor powers. Lake (2009) shows

that states that defer to US security and economic policies have higher volumes of bilateral trade. Machain and Morgan (2013) and Lake (2007) report that increased embeddedness within the US alliance network is associated with a state reducing its defence spending. Moreover, such states are more likely to join US-led international coalitions (Lake, 2009). Johnson (2015) shows that the greater the security risk faced by a state, and the more an ally offsets this risk, the greater the foreign policy concessions of a target to a defender. Minor states are therefore granted greater financial flexibility or security, while their foreign policy options increasingly reflect those of a friendly superpower, such as the US (e.g., Morrow, 1991).

Strategic geo-political concerns colour human rights reports made by the SD. SD reports are important because they affect the international business of firms and the policy decisions of governments and international institutions through naming-and-shaming offenders (and those that interact with them) (Blanton, 2000; Blanton and Blanton, 2007; Lebovic and Voeten, 2009; Murdie and Peksen, 2013; Richards *et al*, 2001). The Human Rights Caucus within the US Congress succeeded in tying foreign aid eligibility to meeting minimal human rights standards. The SD is thus forced to balance geo-political concerns against human rights issues in states that hold strategic military or economic value. SD officials are aware of the political implications that accompany generating negative human rights reports about strategic partners, and may even face pressure to provide favourable ratings.

Figure 3 illustrates these strategic calculations on the part of SD reports. There is little difference in the SD and AI human rights scores for Cuba, a repressive regime that is unfriendly towards and receives no assistance from the US. On the other hand, the SD human rights scores assigned to Israel, a state that traditionally has been close to the US and with a more contested

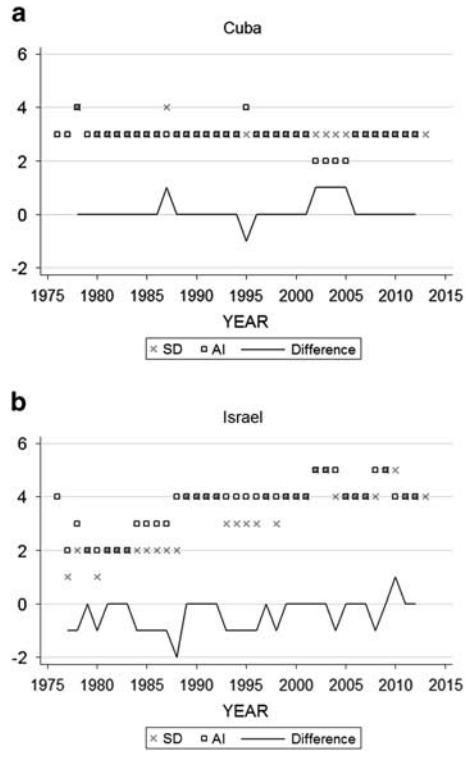


**Figure 3** Comparison of SD and AI scores for friendly and unfriendly regimes.

human rights record, are consistently more favourable than the AI human rights scores, at least before 2000.

The *qualitative* nature of the SD reports are sometimes subtly biased, by, among other points, attributing violence to non-government agents, downplaying the degree of involvement of central government officials, overstating the likelihood of judicial punishment for violators. These qualitative differences in the reports, moreover, may manifest themselves in different quantitative human rights scores when using an objective, pre-established coding scheme, such as those employed by PTS (see Neumann and Graeff, 2013). To demonstrate these differences, we focus on Chile by (i) looking at the AI and SD series of scores, and (ii) examining the content of the reports in 1980.

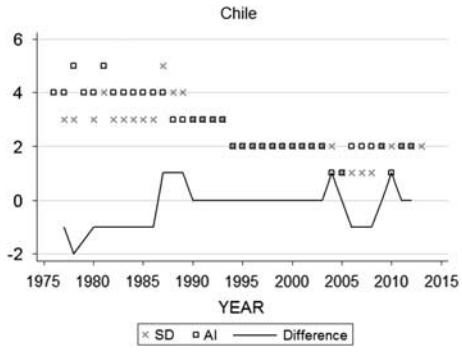


**Figure 4** Comparison of AI and SD scores for Chile.

First, as displayed in Figure 4, from 1977 to 1986 the scores generated by PTS from AI reports were consistently higher (worse rights records) than scores generated from SD reports (notice that there are 2 missing years from the SD reports: 1976, 1979). In 1978, the difference is −2, indicating that the SD is painting a much more favourable picture of its regional ally. Given that the US backed General Augusto Pinochet in the coup of 11 September 1973, this is not terribly surprising. While AI paints a picture in which Chile 1978 and Chile 1981 are coded as having 'terror spread to the whole population … [with] no limits on the means or thoroughness with which [leaders] pursue personal or ideological goals' (i.e., the worst possible human rights conditions), the SD narrows this troubling picture to just the political realm, noting 'extensive political imprisonment … [where] execution or other political murders and brutality may be common … [and] unlimited detention, with or without a trial, for political views is accepted' (quotes from PTS Website).[8] Despite the serious human rights conditions represented by a score of 3, the big difference between the SD and AI reports should cause serious concern for scholars engaging in 'value free' research.

Second, by focusing on particular reports, we can help make more sense of these disparities. We look at AI and SD human rights reports for Chile in 1980.[9] The reports demonstrate anecdotal qualitative evidence of our primary claim – that SD reports are biased towards US 'friends' (and away from US enemies).

The SD report for Chile in 1980 demonstrates the anti-communist ideological orientation of US foreign policy. The leading paragraph concludes that 'the trauma of the Allende period (1970–73) and the view that his policies were leading to a Marxist state continue to influence the attitudes of many Chileans' (Department of State, 1981: 367). The second paragraph of the report begins to detail the serious human rights conditions in Chile, stating that 'in the period 1973–77, the regime undertook to curb dissent through a series of repressive measures, unprecedented in contemporary Chilean history' (ibid.: 367), but the tone has already been set. Despite a long list of grievances ('arbitrary detention and cases of torture', 'basic freedoms of speech and assembly are restricted', and a 'continued "state of emergency" … [giving] the government extraordinary authority similar to that under a state of siege') the SD concludes the third paragraph with the hopeful observation that 'in several instances, however, the courts and press have taken positions defending human rights' (ibid.: 367). Thus, while the US does not necessarily hide human rights violations, it distorts them by minimising their significance. It makes them seem like naturally arising necessities that the Pinochet government had no choice but to engage in to ensure the stability of the country.

The AI report, on the other hand, begins with the claim that 'there was a marked deterioration in the human rights situation during the year' and that the 'major concerns of Amnesty International were political killings, political imprisonment and torture. There were many arbitrary detentions, allegations of torture and harassment of trade unionists, members of youth organisations, human rights activists,

members of church organisations, political opponents, and the poor' (Amnesty International, 1980: 116). The report continues, stating that the 'new constitution, currently being drafted without participation of or consultation with independent lawyers, will have no place for political parties', further noting that independent lawyers have been forbidden from holding meetings (ibid.: 116). The report makes no reference to communism, instead focusing on the state of emergency that the Pinochet government renewed, before providing an exhaustive list of specific violations and allegations culminating in charges that political prisoners were killed and left in unmarked graves (ibid.: 122).

In addition, the tone of the reports differs when discussing the same specific incidents of a human rights violation. Both AI and the SD mention a 1973 incident in Lonquén in which policemen were 'identified as having executed sixteen persons and deposited the bodies in a lime kiln' (US Department of State, 1981: 369). Compare the following descriptions of the continued investigation into the killings:

> *Since 1973 the government has disregarded the interests and feelings of the relatives of disappeared prisoners. In September 1979 the Military Court announced that the bodies found in Lonquén would be returned to their families. The relatives then received the news that their dead had been buried secretly. The Cardinal of Santiago immediately issued a condemnation stating 'human dignity has been violated in the most extreme manner. (Amnesty International, 1980: 121)*

and

> *The special judges are continuing investigations into some cases, have suspended action in numerous others, and have referred still others to military justice when the military or police are believed to have been involved. Some families of the disappeared have appealed the suspension or referral to the military courts. To date, the appeals have had little effect in resolving the whereabouts of their relatives. (US Department of State, 1981: 367)*

There is nothing objectively different in the portrayal of this case. Yet, the picture that emerges from the SD is more forgiving towards the Chilean state: SD reports more generally distinguish the state from its agents, imply that the courts are able to punish wrongdoers, and that the poor human rights situation in Chile is a function in part of violent (Marxist) opposition and state agents who are operating outside of their constitutional prerogative rather than being controlled from the central government. AI, on the other hand, places direct responsibility on the government.

This brief comparison between AI and SD reports for a particular country in a particular year demonstrates anecdotal support that there are *qualitative* differences in the reports. When using a pre-established coding scheme, objective coders could reasonably assign different *quantitative* scores to the same country-year depending on the source.

## TEMPORAL DYNAMICS

Qian and Yanagizawa (2009) and Poe *et al* (2001) both examine the possibility that SD reports may exhibit temporal variation, perhaps varying by president or political party. The head of the SD is one of the top appointees of an administration. It is easy to imagine how ideological directives from the very top could influence the type of language found in SD reports. On the other hand, those who compile reports are not high-level appointees, and they may therefore be immune to presidential politics. As early as the mid-1980s, Neufville (1986: 682) found evidence that reputational costs pressure the SD to

produce a truly unbiased set of documents, which are 'independent of the Administration's political stance'. Poe *et al* (2001) look at Presidential effects, and find differences between the Carter (1977–1981) and Reagan (1981–1989) administrations. However, because of the limited temporal scope, they could not make the obvious step from comparing individual administrations to a more general comparison of parties.

First, it may be reasonable to assume that Republicans are harsher on average than Democrats in their assessment of states' human rights behaviour. The neoconservative foreign policy of George W. Bush (2001–2009) explicitly used human rights rhetoric and pointed to the human rights records of states to justify intervention (Nau, 2013). Thus, a conservative administration would potentially want more critical SD reports to make its enemies seem worse than they actually are. On the other hand, it is also reasonable to assume that Democrats could be harsher on average than Republicans. There are many ideological justifications for intervention based on human rights behaviour, and these are not exclusively associated with one party. In fact, a president's party does not seem to be a good predictor of foreign policy behaviour writ large. Though President Obama (2009–present), for example, tried to change course from President Bush's human rights policies, he failed to make substantive change to US practices (Gibler and Miller, 2012; Roth, 2010). Common wisdom says that fear of being perceived as soft on security issues can pressure Democrats to be even more hawkish than Republicans. Given the forces that could be at work pulling parties in both directions, it seems reasonable to expect a null relationship between $PTS_{AI}$ and $PTS_{SD}$ scores.

It is the null relationship that turns out to be the case in the full sample (1976–2013). Democrats are no harsher than Republicans on average. The mean $PTS_{SD}$ score is 2.38 under Republicans and 2.40 under Democrats, an insignificant difference. However, plotting over time, we see other dynamics that should be explored, namely the convergence between the average SD and AI PTS scores. The SD has historically been much more favourable to countries than AI, but that average difference has shrunk over time. This also plays out on a country-by-country basis.[10] The average difference between SD and AI was greatest in 1977, the first year in which there is data for both AI and SD. Over the course of the Reagan administration, the SD and AI scores converge, with the SD increasing and the AI decreasing. While there continues to be differences with the SD always lower on average than the AI, by the time President Clinton takes office, the average AI and SD scores are not significantly different. This is backed up by the average bias plotted in Figure 5 using the y -axis on the right side ranging from −0.8 to 0.8. In the Clinton administration, there is positive bias (SD is harsher than AI) in 3 years, and a negative bias in 5 years. In the Bush administration (1989–1993), there is positive bias in 5 years, and negative bias in 3 years. For Obama's administration, the first 4 years have all been positive bias.

*'...there are qualitative differences in the reports. When using a pre-established coding scheme, objective coders could reasonably assign different quantitative scores to the same country-year depending on the source'*

There are several possible explanations for this dynamic pattern. The first is that there really has been a change over time –

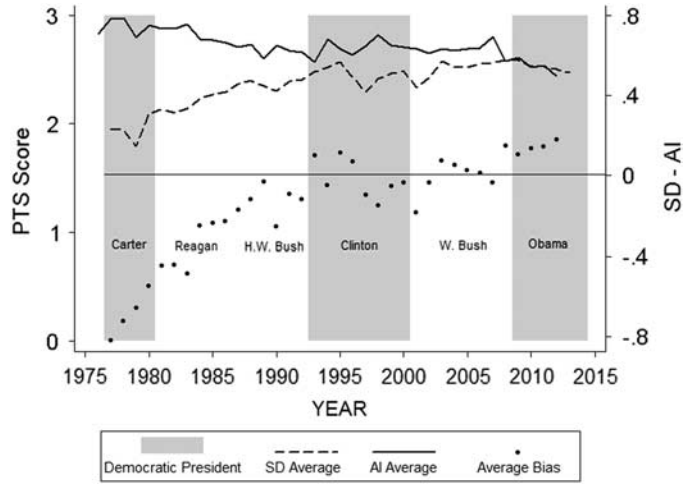


**Figure 5** Average SD and AI PTS scores and SD-AI over time.

human rights have been getting worse. However, this interpretation contradicts recent work, which shows that there is an overall trend that human rights are getting better (Fariss, 2014). Another option is that a structural change in reporting of human rights occurred at the end of the Cold War. The incentives for the SD to paint a prettier (uglier) painting of other states' human rights behaviour radically changed after its competition with the USSR ended.

Finally, the issue may be a measurement one. During the end of the Gerald Ford administration (1974–1977), the legislation was first developed tying the foreign aid decisions of the US to the human rights practices of the recipient states. This policy change is what sparked the SD to start issuing the human rights reports in the first place, and the consequence is that the first years of these reports only contain states that were already receiving US aid or were being considered for it. This biases the sample of countries to only those that are either underdeveloped or held strategic importance in the global competition with the USSR. Only after 1980 did SD reports begin to cover what is now essentially the population of states in the international

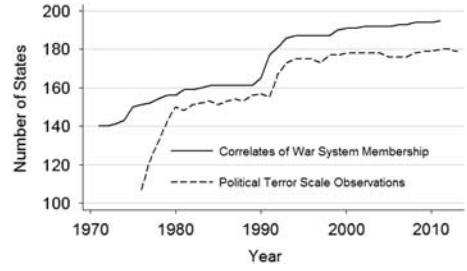


**Figure 6** System membership and political terror scale coverage.

system. Figure 6 compares the coverage by PTS with the number of states in the international system as defined by the Correlates of War (2011).

Poe *et al* (2001: 661) also suggest that 'the overall distribution of the difference variable indicates that the US has tended to be somewhat less harsh than Amnesty in evaluating the human rights practices of other governments'. In the years since their analysis, this result does not seem to hold. While the mean $PTS_{SD}$ score is lower (better human rights practices) until around 2007, it is statistically indistinguishable from $PTS_{AI}$ from around 2003 onwards. The major differences between the two scores are artefacts of the Carter

and Reagan years. In fact, by the time President Bush takes office, the new trend is that for any given country, there does not appear to be a systematic difference in the likelihood of being scored higher or lower based on whether the AI or SD reports are the source for PTS coders. Even though the descriptive statistics tell a story that average bias is disappearing, it may be the case that bias reflecting US strategic interests still occurs despite being masked by general trends.

In the next section, we show that the degree of SD bias is systematic and can be estimated by variables indicating geopolitical value and concern to the US.

## ESTIMATING AND CORRECTING BIAS IN SD SCORES

In this section, we estimate the degree of bias in SD human rights scores. This exercise is helpful for two reasons. First, by demonstrating inherent bias in SD human rights scores, we show that measures based on these reports, that is, CIRI and PTS, are also biased. This is potentially problematic given that these measures are frequently employed within the quantitative human rights literature. Second, because we are able to estimate this bias in a systematic and robust manner, we are able to account and correct for it. The degree of estimated bias can be added to the original SD scores to generate 'corrected' SD scores.[11] This is especially useful given that SD scores are available for a broader spatial sample of states than AI scores.

Our dependent variable is the degree of bias, which we treat as the difference between SD and AI PTS scores, or $bias = PTS_{SD} - PTS_{AI}$. We treat the amount of bias in SD reports as a function of US strategic interests. To capture US strategic interests, we account for the US' global geopolitical conflict with the Soviet Union,

*'...by demonstrating inherent bias in SD human rights scores, we show that measures based on these reports – that is, CIRI and PTS – are also biased'.*

whether a country is a democracy, how embedded it is in the US alliance network, and its relative importance to US trade.[12] We conduct our analysis for the period 1980–2006. We exclude pre-1980 data from this analysis, since the SD did not systematically collect human rights information.

We operationalise *Cold War* as all years prior and up to 1991, when the USSR dissolved. *Democracy* is coded using the 21-point *polity2* scale with greater values indicating more democratic states (Marshall and Jaggers, 2013). *US alliance embeddedness* captures how closely a state is tied to the US in security terms. We follow Lake's (2009) procedure by counting the number of allies that a state shares with the US as a proportion of all of its formal alliances. The logic is that states with non-diversified alliance portfolios are more accepting and central to US foreign policy (i.e., Morrow, 1991). *US alliance embeddedness* is measured as (1)/(*State i's # of Independent Alliances*), where state *i* is assumed to always be allied with itself to avoid undefined values (Lake, 2009: 70, fn 13).[13] Larger values indicate closer US security ties. Finally, *proportion US trade* accounts for the relative economic value of a state from the perspective of the US. *Proportion US trade* is measured as (*US-State i's Bilateral Trade*)/(*Total US trade*), with larger values indicating greater economic importance to the US.

We include cubic polynomials of time to account for the temporal dynamics

reported earlier (Fariss, 2014). Cubic polynomials are more efficient than fixed effects and easier to implement and interpret than cubic splines (Carter and Signorino, 2010). Lastly, we include regional dummies.[14] We do this for two reasons: first, some regions may have special significance to US foreign policy, that is, the Carter Doctrine and the Middle East. Second, the SD is organised into regional offices, which may affect the editing of reports as they are being compiled (Poe *et al*, 2001: 654–655).

## EMPIRICAL RESULTS

Table 1 shows the results of a variety of model specifications used to test the systematic bias in SD reports. We employ two estimation techniques, ordinary least squares (OLS) and ordered logit. For each technique, we employ a minimalist model in which only geo-political variables are included. Next, we include temporal controls in the model, before estimating a model, which includes the previous year's degree of bias. Finally, we estimate a fourth model with regional dummies.

The coefficient for *Cold War* is negative and significant in the first two OLS models. The result indicates that $PTS_{SD}$ is biased downwards during the Cold War. This trend does not hold, however, once a lagged DV is included, nor in any of the ordered logit models, though it is significant at the $p<0.1$ level using a one-tailed test.

The coefficient for *Democracy* is negative and significant in all models. The coefficients associated with *US alliance embeddedness* and *proportion of US trade* are negative across the different models and statistically significant in the minimalist models and those with temporal controls. However, the inclusion of the lagged DV and regional dummies results in a failure for *US alliance embeddedness* to reach statistical significance at traditional levels, though *US alliance embeddedness* is statistically significant at the $p<0.1$ level using a one-tailed test. *Proportion US Trade* is statistically significant even after controlling for the lagged DV, but fails to reach statistical significance once regional dummies are included. Finally, the lagged DV is positive and statistically significant in all models in which it is included.

*Overall, the models reported in Table 1 show consistent support for the claim that State Department reports are biased in favour of countries with strategic value to the United States.*

The net effect of the time trend coefficients is an increase in positive bias. This result makes sense once the size of the constant is taken into account: it is large and negative. The positive time trend reduces the degree of bias over time. Regional dummies show significance relative to the reference category of Western Europe. This means that the degree of bias is greatest in Eastern Europe relative to Western Europe, even controlling for Cold War and other strategic variables.

Overall, the models reported in Table 1 show consistent support for the claim that SD reports are biased in favour of countries with strategic value to the US. To evaluate the strength of our instruments, we report the F-statistic. As a general rule of thumb, an F-statistic greater than 10 indicates that the instrument does not suffer from weak instrument bias (Sovey and Green, 2011; Stock and Watson, 2007). In the OLS models, the F-statistic is largest in the model with the four explanatory variables, the cubic polynomials, and the previous year's bias.

**Table 1: Estimating Bias in $PTS_{SD}$ Scores, 1980–2006**

| Variable | OLS | | | | Ordered Logit | | | |
|---|---|---|---|---|---|---|---|---|
| Cold War | −0.305*** | −0.121* | −0.084 | −0.087 | −0.772*** | −0.263 | −0.190 | −0.200 |
| | (0.030) | (0.066) | (0.066) | (0.066) | (0.076) | (0.169) | (0.176) | (0.176) |
| Democracy | −0.006*** | −0.008*** | −0.006** | −0.005** | −0.015*** | −0.020*** | −0.016*** | −0.015** |
| | (0.002) | (0.002) | (0.002) | (0.003) | (0.006) | (0.006) | (0.006) | (0.007) |
| US Ally Embed. | −0.082** | −0.068* | −0.051 | −0.078 | −0.214** | −0.181* | −0.126 | −0.211 |
| | (0.037) | (0.037) | (0.037) | (0.064) | (0.093) | (0.093) | (0.098) | (0.167) |
| Prop. US Trade | −0.030*** | −0.028*** | −0.027*** | −0.010 | −0.077*** | −0.073*** | −0.064*** | −0.020 |
| | (0.009) | (0.009) | (0.009) | (0.010) | (0.021) | (0.021) | (0.023) | (0.026) |
| Time | | 0.155*** | 0.125*** | 0.125*** | | 0.398*** | 0.320*** | 0.330*** |
| | | (0.031) | (0.032) | (0.032) | | (0.079) | (0.085) | (0.085) |
| $Time^2$ | | −0.008*** | −0.006*** | −0.006*** | | −0.020*** | −0.016*** | −0.017*** |
| | | (0.002) | (0.002) | (0.002) | | (0.005) | (0.006) | (0.006) |
| $Time^3$ | | 0.000*** | 0.000** | 0.000** | | 0.000*** | 0.000** | 0.000** |
| | | (0.000) | (0.000) | (0.000) | | (0.000) | (0.000) | (0.000) |
| Lagged Bias | | | 0.225*** | 0.215*** | | | | |
| | | | (0.017) | (0.017) | | | | |
| Americas | | | | 0.158** | | | | 0.420** |
| | | | | (0.072) | | | | (0.187) |
| East Europe | | | | 0.261*** | | | | 0.692*** |
| | | | | (0.067) | | | | (0.178) |
| Africa | | | | 0.163*** | | | | 0.426*** |
| | | | | (0.055) | | | | (0.145) |
| Middle East | | | | 0.043 | | | | 0.098 |
| | | | | (0.054) | | | | (0.141) |
| Asia | | | | 0.175*** | | | | 0.491*** |
| | | | | (0.068) | | | | (0.180) |
| Oceania | | | | 0.211* | | | | 0.580* |
| | | | | (0.114) | | | | (0.301) |
| Constant | 0.035* | −0.982*** | −0.787*** | −0.914*** | | | | |
| | (0.018) | (0.146) | (0.152) | (0.157) | | | | |
| *F*-Statistic | 38.791 | 29.649 | 48.971 | 30.262 | | | | |
| *R*-squared | 0.045 | 0.059 | 0.113 | 0.121 | | | | |
| Log Likelihood | −3808.554 | −3783.596 | −3408.802 | −3394.309 | −3781.911 | −3754.830 | −3368.408 | −3353.235 |
| Observations | 3,321 | 3,321 | 3,094 | 3,094 | 3,321 | 3,321 | 3,094 | 3,094 |

*Note*: *$p<0.1$, **$p<0.05$, ***$p<0.01$. Standard errors in parentheses. Cut points and lagged binary indicator DVs are suppressed in the ordered logit.

## ADDRESSING BIAS IN SD SCORES

In the previous section, we demonstrated that bias in $PTS_{SD}$ is systematic. We recommend several suggestions to address this. First, we suggest excluding pre-1980 data owing to selection bias. Second, we argue than the degree of estimated bias can be used to 'correct' the original SD scores. The corrected human rights scores can then be subjected to traditional analyses as either the dependent or independent variable.

## EXCLUDING TRUNCATED DATA

Before 1980, the $PTS_{SD}$ sample may be inconsistent with the broader population of states.[15] The lower graph in Figure 7 shows that after 1980, microstates account for most of the missing cases. Before 1980, however, the average population of states not in the sample is quite high. This coincides with the change in SD human rights reporting practices in 1980. The top half of Figure 6 shows another potential source of bias. During the Cold War, non-missing states were more democratic on average than the population of missing states. This result supports the findings reported earlier in Table 1.

These findings support the contention that the data from the 1976–1980 period suffers from a type of selection bias. In the early period, the SD only made human rights reports for current or potential foreign aid recipients (Poe *et al*, 2001). Since foreign aid is correlated with being included in these early reports and with factors related to human rights violations, the data is a truncated sample (Breen, 1996).[16] While it is a sin to omit good data, it is even worse to use bad data (Lewis-Beck, 1995; Breen, 1996). Ignoring this non-representative sample can bias estimates if these early years are included in the larger data set. Currently, most of the states not contained in the PTS data set that are part of the international system by COW are micro-states, those with a very low populations, and researchers are comfortable excluding them without sacrificing validity.[17] As Figure 6 demonstrates, it is only the pre-1980 data that suffers from this problem and ought to be excluded by researchers.

## CONSTRUCTING INSTRUMENTED SD SCORES

Another way to address bias while also potentially expanding the spatial scope of human rights data is to identify and 'correct' for systematic bias. We do this by instrumenting the $PTS_{SD}$ scores. The procedure to instrument $PTS_{SD}$ scores and correct for systematic bias is done in two steps. First, we estimate a model predicting the degree of bias between $PTS_{SD}$ and $PTS_{AI}$. The predicted difference, or $\widehat{bias}$, is estimated from Model 2 of Table 1. Second, we add the value $\widehat{bias}$ to the $PTS_{SD}$ to generate an instrumented variable of PTS, that is, the 'corrected' value of PTS, or $PTS_{IV}$. More formally

$$\widehat{bias} = \beta X + t + t^2 + t^3 \qquad (1)$$

$$PTS_{IV} = PTS_{SD} + \widehat{bias} \qquad (2)$$

where $X$ is a matrix of independent variables, $\beta$ is a vector of coefficients, and $t$ indicates time.

We choose the specification from Model 2 of Table 1 to instrument bias. We select this model because it is parsimonious, theoretically driven, yet still captures temporal dynamics. Additional analyses using other specifications, such as Model 3, which has the largest F-statistic, provide similar results.

Finally, our 'correction' for bias introduces uncertainty into any resulting estimates. To account for this, we

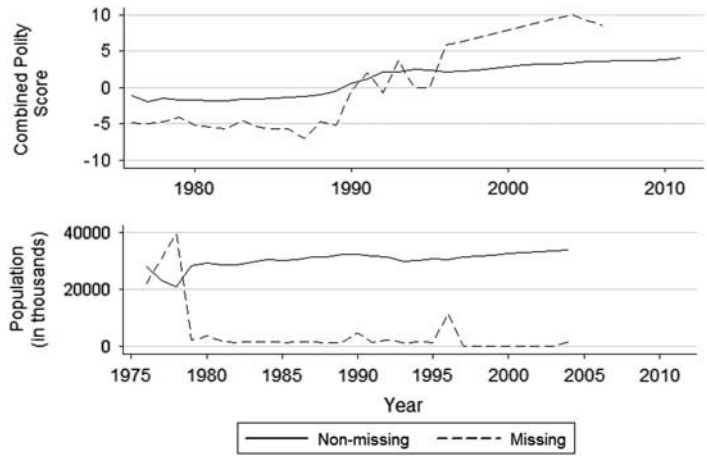


**Figure 7** Comparing the $PTS_{SD}$ sample to the population.

bootstrap 100 replications to calculate the standard errors in the subsequent analysis.

## RE-ESTIMATING REPRESSION AND INTERSTATE CONFLICT

We replicate Wright (2014) to demonstrate the effect of applying instrumented SD scores and dropping pre-1980 data when using $PTS_{SD}$. Wright (2014) argues that states are more likely to engage in domestic repression when they are seeking to revise territory abroad, but that the degree of repression is conditioned by regime type and conflict severity. Democracies are expected to seek territorial revision for intangible qualities – historic or symbolic – that are easy to rally the public behind a s it triggers in-group/out-group dynamics about the nation's identity. Domestic opposition is often viewed as a security threat to the public's security, especially as the severity of the conflict increases (Wright, 2014: 378–379).

Autocratic regimes, on the other hand, focus on delivering private, tangible goods to the elites that maintain the regime. Domestic repression is likely to remain at pre-conflict levels, or even decrease, as security forces are deployed to the front line of the interstate conflict. Any decrease in repression is expected to become more significant as the severity of the interstate conflict increases, as the regime must allocate even more domestic resources abroad (Wright, 2014: 379–380).

Wright tests these expectations by employing an ordered logit for the period 1977–2001 estimating the likelihood of a change in category from the previous year.[18] The dependent variable is $PTS_{AI}$ with $PTS_{SD}$ replacing any missing data, and the primary independent variables are whether a state is revisionist in a territorial conflict (*territorial revision*), the severity of the conflict (*MID fatalities*), and their interaction (*terr. rev. X fatalities*). Wright runs analyses on the full sample, as well as running additional analyses on subsamples of democratic and autocratic states. We refer the reader to Wright (2014: 380–382) for an in-depth discussion and the sources of all independent and control variables.

Table 2 presents the result of the exact replication of Wright (2014, Table 1: Models 2, 4, and 6), and a replication excluding the pre-1980 data. Table 3 reports a replication for the full time period 1976–2001 reported in Wright (2014) using our instrument $PTS_{IV}$ in the

cases were $PTS_{SD}$ is used to fill in missing data (approximately 14 per cent of the data in the full sample)[19] and finally, a replication using our proposed best practice of excluding the pre-1980 data and employing PTS in the place of PTSSD for missing data.[20] We restrict our replication to include only observations used in the original analysis.

Looking at Table 2, the results of the exact replication of Wright shows that for the full sample, only *territorial revision* is statistically significant. The democratic and autocratic sub-samples, however, report the interaction term *terr. rev* × *fatalities* to be statistically significant and in the predicted direction for each regime type. Moreover, the constitutive term *MID fatalities* is negative and significant in the democracy sub-sample, while both *territorial revision* and *MID fatalities* are positive and significant in the autocracy sub-sample.[21]

In the replication excluding the truncated pre-1980 data, the interaction term *terr. rev.* × *fatalities* is not statistically significant in any model, though it does approach traditional significant levels ($p>0.104$) in the democratic sub-sample. The constitutive term *MID fatalities* is positive and statistically significant in the autocracy sub-sample, while the other constitutive terms fail to reach statistical significance. These changes are consistent with selection bias in the pre-1980 data.

Table 3 utilises the instrument approach described in the previous section. The first set of models apply our instrumental variable, $PTS_{IV}$, in place of the $PTS_{SD}$ values used for repression data when AI scores are unavailable for the full time period. As was the case in the original analysis, *territorial revision* is a statistically significant variable in the full model, while neither *MID fatalities* nor their interaction term are.[22] Likewise, the interaction term and the constitutive terms are significant in the autocracy sub-sample, and in the same direction as the original analysis, but are not significant in the democracy sub-sample.

Next, we follow our proposed best practice of both excluding the pre-1980 data and using $PTS_{IV}$ in place of $PTS_{SD}$ scores. In this case, *MID fatalities* is positive and statistically significant in the autocracy sub-sample, while neither the interaction term nor the other constitutive term is statistically significant, nor approach traditional significance levels, in any of the samples. Despite the change in findings concerning the interaction and constitutive variables, other variables are consistent with the original analyses across the model specifications.

In the models that apply either of our suggested corrections, the substantive inferences change compared with the original analyses. Once the systematic bias in $PTS_{SD}$ is accounted for, *territorial revision, MID fatalities* and *terr. rev* × *fatalities* are no longer statistically significant in the democracy sub-sample. Similarly, after excluding the selection bias in the pre-1980 data, only *MID fatalities* in the autocracy sub-sample is statistically significant among the interaction and constitutive variables. These changes highlight the importance of accounting for both types of bias in PTS scores.

## CONCLUSION

We laud recent attempts to widen the scope and improve the validity of human rights data (e.g., Schnakenberg and Fariss, 2014). In general, improved methodological sophistication has been a trademark of quantitative human rights research in the last several years, whether through innovative use of existing data (Hill and Jones, 2014; Schnakenberg and Fariss, 2014; Fariss, 2014) or by finding new sources of data to push human rights research into new territory (Conrad *et al*,

**Table 2: Territorial Revision and State Repression: Exact Replication and Exclusion of Pre-1980 Data**

| | Exact Replication | | | Excluding Pre-1980 | | |
|---|---|---|---|---|---|---|
| | All | Dem | Autoc | All | Dem | Autoc |
| Territorial Revision | 0.354** | −0.154 | 0.506** | 0.188 | −0.159 | 0.288 |
| | (0.162) | (0.257) | (0.232) | (0.176) | (0.283) | (0.233) |
| MID Fatalities | 0.042 | −0.182* | 0.115* | 0.042 | −0.175 | 0.124* |
| | (0.060) | (0.102) | (0.066) | (0.061) | (0.130) | (0.072) |
| Terr. Rev. | −0.077 | 0.329** | −0.179* | −0.040 | 0.318 | −0.154 |
| × Fatalities | (0.090) | (0.134) | (0.108) | (0.094) | (0.196) | (0.110) |
| Population | 0.136*** | 0.162*** | 0.163*** | 0.143*** | 0.154*** | 0.182*** |
| | (0.034) | (0.062) | (0.040) | (0.028) | (0.049) | (0.035) |
| Econ. Development | −0.211*** | −0.394*** | −0.111* | −0.234*** | −0.412*** | −0.131** |
| | (0.046) | (0.090) | (0.061) | (0.043) | (0.081) | (0.052) |
| Civil Conflict | 1.429*** | 1.573*** | 1.421*** | 1.478*** | 1.668*** | 1.467*** |
| | (0.157) | (0.368) | (0.152) | (0.119) | (0.251) | (0.138) |
| Non-fatal MID | 0.128 | 0.003 | 0.284*** | 0.166* | 0.067 | 0.306*** |
| | (0.085) | (0.149) | (0.104) | (0.091) | (0.153) | (0.115) |
| Democracy | −0.616*** | | | −0.556* | | |
| | (0.114) | | | (0.095) | | |
| Log-Likelihood | −3065.023 | −981.538 | −2038.022 | −2609.538 | −879.994 | −1691.004 |
| Observations | 3,538 | 1,353 | 2,185 | 3,034 | 1,217 | 1,817 |

*Note*: $*$ $p<0.1$, $**$ $p<0.05$, $***$ $p<0.01$. Replication of Wright (2014). Standard errors in parentheses and clustered by country. Cut points and lagged dependent variables are calculated but not displayed.

**Table 3: Territorial Revision and State Repression: Using Instrumental Variables**

| | $PTS_{IV}$ | | | Excluding Pre-1980 and use $PTS_{IV}$ | | |
|---|---|---|---|---|---|---|
| | All | Dem | Autoc | All | Dem | Autoc |
| Territorial Revision | 0.275* | −0.250 | 0.428** | 0.103 | −0.281 | 0.194 |
| | (0.150) | (0.273) | (0.208) | (0.174) | (0.306) | (0.258) |
| MID Fatalities | 0.046 | −0.155 | 0.117** | 0.052 | −0.166 | 0.139* |
| | (0.060) | (0.150) | (0.058) | (0.062) | (0.160) | (0.073) |
| Terr. Rev. | −0.060 | 0.376 | −0.171* | −0.022 | 0.397 | −0.152 |
| × Fatalities | (0.092) | (0.242) | (0.097) | (0.104) | (0.248) | (0.129) |
| Population | 0.132*** | 0.159*** | 0.170*** | 0.150*** | 0.165*** | 0.201*** |
| | (0.027) | (0.043) | (0.031) | (0.024) | (0.044) | (0.038) |
| Econ. Development | −0.224*** | −0.427*** | −0.110*** | −0.247*** | −0.439*** | −0.129*** |
| | (0.038) | (0.071) | (0.045) | (0.036) | (0.088) | (0.048) |
| Civil Conflict | 1.384*** | 1.476*** | 1.386*** | 1.459*** | 1.624*** | 1.457*** |
| | (0.108) | (0.0276) | (0.135) | (0.139) | (0.310) | (0.143) |
| Non−fatal MID | 0.124* | −0.033 | 0.320*** | 0.187** | 0.066 | 0.368*** |
| | (0.072) | (0.123) | (0.097) | (0.092) | (0.147) | (0.102) |
| Democracy | −0.624*** | | | −0.596*** | | |
| | (0.083) | | | (0.099) | | |
| Log-Likelihood | −2978.654 | −961.498 | −1963.649 | −2623.155 | −888.195 | −1686.261 |
| Observations | 3,361 | 1,279 | 2,082 | 2,998 | 1,207 | 1,791 |

*Note*: $*$ $p<0.1$, $**$ $p<0.05$, $***$ $p<0.01$. Replication of Wright (2014). Standard errors calculated from bootstrap of 100 replications. Cut points and lagged dependent variables are calculated but not displayed.

2014; Murdie and Davis, 2012). Yet, it is nonetheless the case that the field of human rights research still fundamentally rely on expert coded indexes drawn from SD and AI reports. As we have shown, any study using measures constructed from SD reports, such as the important CIRI and PTS indicators, is susceptible to bias reflecting US geo-political concerns, especially the further back in time one goes. To reiterate, if the data used in the construction of a measure are biased, the measure itself will be biased (Neumann and Graeff, 2013). This is substantively important because the most widely used indexes in the quantitative study of human rights – CIRI and PTS – use those reports. This bias is somewhat diffused, but not eliminated, by including these measures as part of a larger index (e.g., Schnakenberg and Fariss, 2014), or when averaging two scores (i.e., average of $PTS_{AI}$ and $PTS_{SD}$). Ignoring this bias is a serious problem because it can change the inferences that we make from quantitative analysis of human rights data, as illustrated in our replication.

This does not mean, however, that quantitative human rights researchers should abandon their fight to shed greater light on the human rights behaviour of states. On the contrary, this task is an important one for academics who wish to understand the phenomenon from a scientific standpoint, but who are also interested in making the world a less repressive place. Our primary concern is to warn researchers of the pitfalls of uncritically employing measures, which are introducing bias into their analyses.

We offer some solutions to address the bias in SD reports. The first is simple: if there are known biases in the data, the data should be excluded. While we found evidence that the full 1976–2013 time period expresses unwanted bias, we are especially worried about the pre-1980 period. The truncated nature of the data is well known, yet the 1976–1980 period is still commonly included with the more systematically collected post-1980 data. Our second suggestion is to explicitly account for the bias by instrumenting it and subtracting it from raw SD scores. As long as appropriate independent variables are available, the expected difference between SD and AI scores can be estimated to generate unbiased data. This is especially useful since SD data are generally available for a wider regional scope than AI scores. Rather than the common practice of replacing the missing AI score with a possibly biased SD score, we suggest using the estimated difference between SD and AI from the subset of data where AI are available to create an unbiased human rights score with the maximum spatial scope.

## Notes

1 Earlier work has dealt with the potential bias included in SD reports. The construction of the CIRI Physical Integrity Rights index takes that bias for granted. The index starts with SD reports, and then coders also use AI's Annual Report. When there is a difference between the two sources, 'our coders treat the Amnesty International assessment as authoritative. Most scholars believe that this step, crosschecking the Country Reports assessment against the Amnesty International assessment, is necessary to remove a potential bias in favour of US allies' (Cingranelli and Richards, 2010: 400). Alternatively, Hill *et al* (2013) show that while some NGOs face their own strategic interests to inflate allegations of government abuse, AI strictly adheres to their credibility criterion in human rights reporting.
2 For instance, Nordås and Davenport (2013) finds support that an important demographic variable, youth bulges, leads to worse human rights practice. This new variable had not been considered until recently, but it has found robust support. Other recent additions to the cannon of covariates include variables measuring NGO shaming (Ron *et al*, 2005; Murdie and Davis, 2012), international legal instruments (with a focus on the International Covenant on Civil and Political Rights (ICCPR) and the Convention Against Torture (CAT) (see Simmons (2009), Hafner-Burton and Tsutsui (2007), and Hill

(2010) for an introduction to the debate on the role of human rights treaties, and Mitchell *et al* (2013) for domestic legal traditions. See Hill and Jones (2014) for a review of the empirical human rights literature).

3 While we focus on PTS because it disaggregates the data by AI and SD, the CIRI data may have the same set of problems. Unfortunately, we cannot test this assertion directly because of the differing ways that CIRI and PTS use AI and SD reports to construct their indexes. The CIRI Physical Integrity Rights index starts with SD reports, then uses AI reports (where available) to corroborate the scores. Whenever the reports produce different scores, the scores generated from AI reports are used because of the potential biased nature of the SD reports. This attempt to account for SD bias is good in theory, but it is also why CIRI remains subject to the criticisms developed in this article. While CIRI is 'fixing' all the countries for which there are both SD and AI reports, it is ignoring countries for which there are no AI reports. The result is a data set in which there is a split population – those states for which there are both SD and AI reports (resulting in unbiased scores) and those for which there are only SD scores (resulting in potentially biased scores). If there is no relationship between whether AI reports on a country and important covariates, this will simply increase the 'noise' associated with estimation. On the other hand, if where AI does its reporting is associated with covariates, then researchers may be including biased scores into their analyses. AI is less likely to report on countries that the SD includes for non-random reasons. For instance, AI is less likely to include smaller states and states with high GDP/capita. Both economic and demographic variables are important in theories of human rights behaviour, and scholars will be keen to include such variables in their analyses. Since those variables are correlated with whether AI covers them or not, CIRI will suffer from the same bias we have identified in the PTS data, if only to a lesser degree.

4 PTS is not just the human rights conditions writ large; it specifically relates to a subset of rights called personal/physical integrity rights. Moreover, the measure is limited to the government's behaviour with respect to those rights. For instance, a country is not held responsible for human rights abuses committed by rebel groups who are operating in the nominal territory of the state. On the other hand, acts of police, even though they may not be condoned at the highest levels of the government, are seen as acts of state despite the fact that those abuses may result from an inability of the state to control its own agents.

5 While it is possible for the $PTS_{SD}$ and $PTS_{AI}$ to differ by four points, in practice, the largest difference is three points.

6 Poe *et al* (2001) subtract the $PTS_{SD}$ from the $PTS_{AI}$, while we do the opposite. If readers are comparing our results to theirs, they will need to flip signs, but the substantive findings will remain unaltered.

7 The fact that zero is the most common category is consistent with Poe *et al* (2001): 54.3 per cent in the full sample and 54.7 per cent in Poe's sample. However, the weighted centre of the distribution is closer to zero in the full sample ($\mu_{SD_i-AI} = -0.09$) compared to Poe, Carey and Vazquez's sample (−0.21), which indicates that the distribution has become more normal over time.

8 The quotations are taken from the descriptions of the different values of the ordinal scale as discussed on the PTS website: http://www.politicalterrorscale.org/ptsdata.php, accessed 5 September 2014.

9 We obtained the report from Amnesty directly from their website: https://www.amnesty.org/en/search/?documentType=Annual+Report&sort=date&p=6, accessed September 15, 2014. The State Department report is archived by the Hathi Trust and is available at: https://babel.hathitrust.org/cgi/pt?id=mdp.39015014143476;view=1up;seq=379, accessed 15 September 2014.

10 Note that the black square markers do not indicate the difference of the average $\mu_{SD}-\mu_{AI}$, but the average of the difference $\mu_{SD-AI}$. In Figure 2 from (Poe *et al*, 2001: 662), the authors report $\mu_{SD}-\mu_{AI}$ over time even though this is not the concept they discuss throughout the rest of the article. Thus, our analysis is not an exact replication/update of their work.

11 Corrected values can be treated as continuous or censored to create an ordinal scale, reflecting the ordinal scale used with PTS.

12 It is worth noting that we do not contend that variables such as trade have no impact on observed human rights performance. Instead, we argue that states holding strategic interest to the US are more likely to have favourable SD reports. That variables like trade may improve actual human rights performance and that it results in more biased SD reports are not mutually exclusive. Any improvement in actual human rights performance should be reflected in both the AI and SD reports. Where these reports diverge, however, we can use the difference between them to predict bias. Moreover, it is not surprising that SD reports reflect US strategic goals. The US is known to consider its strategic interests when creating reports and forecasts. Sahin (2014), for instance, finds that International Monetary Fund (IMF) country forecasts include high degrees of politically motivated bias, reflecting US commitments rather than economic fundamentals. The same US strategic interests have been found to influence decision regarding lending decisions (Barro and Lee, 2005; Stone, 2002, 2004) and foreign aid allocations (Alesina and Dollar, 2000; Bearce and Tirone, 2010; Fleck and Kilby, 2010).

13 We treat states as 'allies' if they are coded as sharing a neutrality, non-aggression, or defensive pact by the Correlates of War alliance data set (Gibler, 2009).
14 We code regions in the following manner: states with Correlates of War country codes between 1–199 are coded as Americas. Country codes 200–399 are coded as *Western Europe*, except for former Communist states that succeeded the USSR or Yugoslavia, or that had been part of the Warsaw Pact, who are coded as *Eastern Europe*. Country codes 400–599 are coded as *Africa*, 600–799 are coded as *Middle East*, 800–899 are coded as *Asia*, and 900–999 are coded as *Oceania*.
15 The time period covered in the CIRI data set starts after 1980 to reflect this problem. Our analysis supports the CIRI project's decision to limit the temporal scope of their data set.
16 Breen (1996: 4) defines the particular type of truncation that concerns this case as a 'sample selected:' '*y* is observed only if some criterion defined in terms of another random variable, *z* is met, such as if $z = 1$'.
17 For many quantitative studies, it is common to purposefully restrict the sample to states with a population of at least 100,000.
18 Following Wright, we include a series of lagged binary variables of the categories of the dependent variable to account for temporal auto-correlation.
19 We calculate predicted pre-1980 $PTS_{IV}$ scores using the values from Model 2 of Table 1.
20 We use cut points of 0.5 to recode these instrumented PTS scores into the ordinal scale used by Wright (2014), that is, scores $<1.5$ are coded as 1, [1.5, 2.5] are coded as 2, [2.5, 3.5] are coded as 3, [3.5, 4.5] are coded as 4, and scores $\geq 4.5$ are coded as 5. We run additional analyses on the unscaled instrumented PTS scores using OLS. Substantive results are the same.'
21 Constitutive terms each have meaningful zero values, enabling their direct interpretation. Testing for the statistical significance of the interaction term requires calculating the covariance of the interaction and constitutive terms, however, making direct interpretation more difficult. Graphical outputs of the marginal effects support the results discussed above.
22 The reported results are similar even if we do not bootstrap the standard errors.

## References


Alesina, A. and Dollar, D. (2000) 'Who gives foreign aid to whom and why?' *Journal of Economic Growth* 5(1): 33–63.

Amnesty International. (1980) *Amnesty International Report 1980*, London: Amnesty International Publications.

Barro, R.J. and Lee, J.-W. (2005) 'IMF programs: Who is chosen and what are the effects?' *Journal of Monetary Economics* 52(7): 1245–1269.

Bearce, D.H. and Tirone, D.C. (2010) 'Foreign aid effectiveness and the strategic goals of donor governments', *Journal of Politics* 72(3): 837–851.

Blanton, S.L. (2000) 'Promoting human rights and democracy in the developing world: US rhetoric versus US arms exports', *American Journal of Political Science* 44(1): 123–131.

Blanton, S.L. and Blanton, R.G. (2007) 'What attracts foreign investors? An examination of human rights and foreign direct investment', *Journal of Politics* 69(1): 143–155.

Breen, R. (1996) *Regression Models: Censored, Sample Selected, Or Truncated Data*, Thousand Oaks, CA: SAGE.

Carter, D.B. and Signorino, C.S. (2010) 'Back to the future: Modeling time dependence in binary data', *Political Analysis* 18(3): 271–292.

Cingranelli, D.L. and Richards, D.L. (2010) 'The Cingranelli and Richards (CIRI) human rights data project', *Human Rights Quarterly* 32(2): 401–424.

Conrad, C.R., Haglund, J. and Moore, W.H. (2014) 'Torture allegations as events data: Introducing the ill-treatment and torture (ITT) specific allegation data', *Journal of Peace Research* 51(3): 429–438.

Correlates of War Project. (2011) 'State system membership list, v2011', available at http://correlatesofwar.org, accessed 1 September 2014.

Department of State. (1981) *Country Reports on Human Rights Practices*, Washington DC: U.S. Government Printing Office.

Esaray, J. and Danneman, N. (2014) 'A quantitative method for substantive robustness assessment', *Political Science Research and Methods* 3(1): 95–111.

Fariss, C.J. (2014) 'Respect for human rights has improved over time: Modeling the changing standard of accountability', *American Political Science Review* 108(2): 297–318.

Fleck, R.K. and Kilby, C. (2010) 'Changing aid regimes? U.S. foreign aid from the cold war to the war on terror', *Journal of Development Economics* 91(2): 185–197.
Gibler, D.M. (2009) *International Military Alliances, 1648–2008*, Washington DC: CQ Press.
Gibler, D.M. and Miller, S.V. (2012) 'Comparing the foreign aid policies of Presidents Bush and Obama', *Social Science Quarterly* 93(5): 1202–1217.
Gibney, M. and Dalton, M. (1996) 'The political terror scale', *Policy Studies and Developing Nations* 4(1): 73–84.
Hafner-Burton, E.M. and Tsutsui, K. (2007) 'Justice lost! The failure of international human rights law to matter where needed most', *Journal of Peace Research* 44(4): 407–425.
Hill, D.W. (2010) 'Estimating the effects of human rights treaties on state behavior', *Journal of Politics* 72(4): 1161–1174.
Hill, D.W. and Jones, Z.M. (2014) 'An empirical evaluation of explanations for state repression', *American Political Science Review* 108(3): 661–687.
Hill, D.W., Moore, W.H. and Mukherjee, B. (2013) 'Information politics versus organizational incentives: When are amnesty international's 'naming and shaming' reports biased?' *International Studies Quarterly* 57(2): 219–232.
Johnson, J.C. (2015) 'The cost of security: Foreign policy concessions and military alliances', *Journal of Peace Research* 52(5): 665–679.
Kalyvas, S.N. and Balcells, L. (2010) 'International system and technologies of rebellion: How the end of the cold war shaped internal conflict', *American Political Science Review* 104(3): 415–429.
Lake, D.A. (2007) 'Escape from the state of nature', *International Security* 32(1): 47–79.
Lake, D.A. (2009) *Hierarchy in International Relations*, Ithaca, NY: Cornell University Press.
Lebovic, J.H. and Voeten, E. (2009) 'The cost of shame: International organizations and foreign aid in the punishing of human rights violators', *Journal of Peace Research* 46(1): 79–97.
Lewis-Beck, M.S. (1995) *Data Analysis: An Introduction*, Thousand Oaks, CA: SAGE.
Machain, C.M. and Morgan, T.C. (2013) 'The effect of US troop deployment on host states foreign policy', *Armed Forces & Society* 39(1): 102–123.
Marshall, M.G. and Jaggers, K. (2013) 'Polity IV Project: Political Regime Characteristics and Transitions, 1800–2013, Version p4v2013e [Computer File]'.
Meernik, J., Krueger, E.L. and Poe, S.C. (1998) 'Testing models of U.S. foreign policy: Foreign aid during and after the cold war', *Journal of Politics* 60(1): 63–85.
Mitchell, S.M., Ring, J.J. and Spellman, M.K. (2013) 'Domestic legal traditions and states' human rights practices', *Journal of Peace Research* 50(2): 189–202.
Morrow, J. (1991) 'Alliances and asymmetry: An alternative to the capability aggregation model of alliances', *American Journal of Political Science* 35(4): 904–933.
Murdie, A. and Davis, D.R. (2012) 'Shaming and blaming: Using events data to assess the impact of human rights INGOs', *International Studies Quarterly* 56(1): 1–16.
Murdie, A. and Peksen, D. (2013) 'The impact of human rights INGO activities on economic sanctions', *Review of International Organizations* 8(1): 33–53.
Nau, H.R. (2013) *Conservative Internationalism: Armed Diplomacy Under Jefferson, Polk, Truman, and Reagan*, Princeton, NJ: Princeton University Press.
Neufville, J.I. (1986) 'Human rights reporting as a policy tool: An examination of the state department country reports', *Human Rights Quarterly* 8(4): 681–699.
Neumann, R. and Graeff, P. (2013) 'Method bias in comparative research: Problems of construct validity as exemplified by the measurement of ethnic diversity', *Journal of Mathematical Sociology* 37(2): 85–112.
Nordås, R. and Davenport, C. (2013) 'Fight the youth: Youth bulges and state repression', *American Journal of Political Science* 57(4): 926–940.
Poe, S.C., Carey, S.C. and Vazquez, T.C. (2001) 'How are these pictures different? A quantitative comparison of the US state department and amnesty international human rights reports, 1976–1995', *Human Rights Quarterly* 23(3): 650–677.
Poe, S.C. and Meernik, J. (1995) 'U.S. military aid in the eighties: A global analysis', *Journal of Peace Research* 32(4): 399–412.
Qian, N. and Yanagizawa, D. (2009) 'The strategic determinants of U.S. human rights reporting: Evidence from the cold war', *Journal of the European Economic Association* 7(2–3): 446–457.
Richards, D.L., Gelleny, R.D. and Sacko, D.H. (2001) 'Money with a mean streak? Foreign economic penetration and government respect for human rights in developing countries', *International Studies Quarterly* 45(2): 219–239.

Ron, J., Ramos, H. and Rodgers, K. (2005) 'Transnational information politics: NGO Human rights reporting, 1986–2000', *International Studies Quarterly* 49(3): 557–588.
Roth, K. (2010) 'Empty promises? Obama's hesitant embrace of human rights', *Foreign Affairs* 89(2): c10–c16.
Sahin, A. (2014) International organizations as information providers: How investors and governments utilize optimistic IMF forecasts. Doctoral Dissertation. Washington University in St Louis. Retrieved from Ann Arbor, MI: Proquest/UMI.
Schnakenberg, K.E. and Fariss, C.J. (2014) 'Dynamic patterns of human rights practices', *Political Science Research and Methods* 2(1): 1–31.
Simmons, B.A. (2009) *Mobilizing for Human Rights: International Law in Domestic Politics*, Cambridge: Cambridge University Press.
Sovey, A.J. and Green, D.P. (2011) 'Instrumental variables estimation in political science: A readers' guide', *American Journal of Political Science* 55(1): 188–200.
Stock, J.H. and Watson, M.W. (2007) *Introduction to Econometrics: International Edition*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.
Stone, R.W. (2002) *Lending Credibility: The International Monetary Fund and the Post-Communist Transition*, Princeton, NJ: Princeton University Press.
Stone, R.W. (2004) 'The political economy of IMF lending in Africa', *American Political Science Review* 98(4): 577–591.
US Department of State. (1981) Country reports on human rights practices, available at https://babel.hathitrust.org/cgi/pt? id=mdp.39015014143476;view=1up;seq=379, accessed 15 September 2014.
US Department of State. (2012) 'Foreign Affairs Manual.' Personnel: Assignments and Details.
Westad, O.A. (1992) 'Rethinking revolutions: The cold war in the third world', *Journal of Conflict Resolution* 29(4): 455–464.
Westad, O.A. (2005) *The Global Cold War: Third World Interventions and the Making of Our Times*, Cambridge: Cambridge University Press.
Wood, R.M. and Gibney, M. (2010) 'The political terror scale (PTS): A re-introduction and a comparison to CIRI', *Human Rights Quarterly* 32(2): 367–400.
Wright, T.M. (2014) 'Territorial revision and state repression', *Journal of Peace Research* 51(3): 375–387.

## About the Authors

**Mark David Nieman** is Assistant Professor of Political Science at Iowa State University. His research interests include conflict processes, state development, and estimating strategic interactions. He has recently published on these topics in *Political Analysis, Conflict Management and Peace Science, Political Science Research and Methods*, and *International Interactions*.

**Jonathan J. Ring** is a Postdoctoral Research Fellow at the University of Michigan. His research focuses on international norm dynamics and human rights, and his recent work has appeared in the *Journal of Peace Research*.