Appendix for

Conrad, Haglund & Moore (2014)

"Torture Allegations as Events Data: Introducing the Ill-Treatment and Torture (ITT) Specific Allegations Data"

Journal of Peace Research

1 Introduction

In this Appendix we provide additional information about the ITT SA data, and elbaorate on several issues we raise in the article. We begin by describing the recruitment and training of the ITT coders, and then turn to the issue of the reliability of the data. A discussion of the statsitical modeling issues raised by using content analysis data from the naming and shaming of a human rights watch dog to study states' (lack of) respect for human rights (in our case, the CAT) follows. We then include a number of additional descriptive figures, without comment, that readers may find of interest.

We defined torture according to the United Nations Convention Against Torture (CAT):

> torture means any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or intimidating or coercing him or a third person, or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity...

When AI uses the terms "ill treatment," "mental abuse," or "physical abuse," ITT codes it as an allegation of ill-treatment (Conrad & Moore 2010a, pp. 28-9).

2 Additional Information about the ITT SA Data

2.1 Description of ITT Specific Allegation (SA) Variables

Because it is event data, the ITT SA data includes information on characteristics of torture allegations typically absent from country-year data on rights violations. Aside from information on the type of victim tortured, the state agent responsible, and the alleged type of torture, the ITT SA data codes variables on individual and government response to allegations, including whether AI indicates that an individual filed a formal complaint, whether or not an investigation was conducted, and the outcome of investigation. Below we describe the key variables in the ITT SA data before turning to more specific descriptive analyses.[1]

[1] Data on Victim Type (VT) and Agency of Control (AoC) are also included in the ITT CY data at the country-year unit of observation (Conrad et al. 2013a).

Order of Magnitude is an ordinal indicator of the number of victims tortured in a specific allegation.[2]

Turning to **Victim Type (VT)**, the ITT SA data distinguish among four economic, social, and/or political groups: Criminal, Political Dissident, Member of a Marginalized Group, and State Agent (Conrad & Moore 2011, 9). Political dissidents include prisoners of conscience, human rights defenders, and protestors. Marginalized individuals include immigrants, members of marginalized ethnic or religious groups, and the elderly or youths. Values on VT are not mutually exclusive, as some victims may exhibit more than one identity in our classification scheme.

Agency of Control (AoC) is an indicator of the domestic institution alleged to be responsible for a given torture or ill-treatment violation. ITT coders distinguished between six government agencies: Police, Prison, Military, Intelligence, Immigration Detention, and Paramilitary (Conrad & Moore 2011, 12–13). Values on AoC are not mutually exclusive, as AI sometimes alleges that a victim is abused by more than one agency.

AI sometimes issues statements of "official concern" that a person is at grave risk to torture or ill-treatment. Other reports indicate that AI *believes* torture occurred in the past, but is not certain about the allegation. To distinguish these allegations from allegations for which AI expresses certainty about a violation, ITT codes the binary **Expectation of Torture** variable as 1 (Conrad & Moore 2011, 10).

The ITT SA Data also provide information on the type(s) of torture alleged in AI documents including: **Ill-treatment, Unstated Torture, Scarring Torture, and Stealth Torture**. Each of these is a binary variable indicating whether AI made an allegation of a particular type of abuse. Torture types are not mutually exclusive. **Scarring Torture** is coded when AI alleges torture that leaves marks on the human body (Conrad & Moore 2011, 11-12), and **Stealth Torture** or "clean" torture is coded for allegations that do not leave marks on the body (Rejali 2007). **Unstated Torture** distinguishes allegations of torture in which AI documents that torture occurred, but does not provide information regarding the type of torture alleged. The CAT not only prohibits torture, but also proscribes states from engaging in cruel, inhuman, or degrading treatment or punishment. **Ill-Treatment** is coded when AI alleges such behavior (Conrad & Moore 2011, 11–12).

The SA data include a number of variables that indicate individual and state responses to torture. **Formal Complaint** indicates whether an allegation of torture was formally reported to the state by either the victim or by NGOs and other like-groups. Formal complaint is coded on a trichotomous scale, where 1 indicates that a formal complaint was filed, 0 indicates that it is not known whether a formal complaint was filed, and -1 indicates that AI specifically mentioned that no formal complaint was filed. **Investigation** indicates whether an allegation of torture was investigated by state authorities (Conrad & Moore 2011, 14) and is coded on the same trichotomous scale as the formal complaint variable. If an in-

[2] When AI provides information on the specific number of victims, coders also recorded the integer value. Note that events in the ITT SA data can be against any positive integer number of of victims. For some research questions, the number of victims per AI allegation may not be relevant. For other questions, researchers may consider "weighing" allegations by the order of magnitude or the integer number of victims.

vestigation occurs, **Investigation Outcome** is coded. This variable indicates whether an adjudication/mediation procedure, administrative sanction, termination of employment, or the creation of legislation followed the investigation (Conrad & Moore 2011, 14). In allegations where adjudication or mediation procedures took place, ITT coders coded **Location of Adjudication**, which indicates whether proceedings took place domestically or internationally. Finally, for any given allegation where adjudication or mediation took place, **Outcome of Adjudication** is coded, which reports whether adjudication resulted in a pardon, conviction or plea, acquittal, or compensation (Conrad & Moore 2011, 15).

The CAT requires that no state expel, return, refoul, or extradite a person to another state where an individual is likely to be in danger of being tortured. **Transborder Torture** is coded when AI makes an allegation of refoulment or extradition (Conrad & Moore 2011, 16). The *expectation* of refoulment or extradition is not coded; refoulment or extradition must have actually occurred for this variable to be coded. In allegations of trans-border torture, the **Destination** of trans-border torture is also coded, indicating the state to which an individual or group of individuals were sent to be tortured.

3 AI Allegations & Human Rights Data Collection

3.1 AI Processes

Several well-known human rights data collection efforts also code data based on US State Department Reports (e.g., Gibney & Dalton 1996, Cingranelli & Richards 2010). The ITT data exclusively codes AI reports for two reasons. First, US State Department reports focus on broad trends in human rights rather then reporting individual allegations of torture. Second, we have conducted interviews with AI staffers to determine the process by which allegations and reports are generated at AI. Such information allows us to better model the data generating process.

AI documents are particularly useful sources because the organization has an extensive quality control procedure that includes research teams composed of both subject and area experts as well as approval by a system of review and approval among AI researchers. The *Annual Report* has the most extensive process vetting procedure, but all AI documents require at least two independent sources of information prior to publishing. Furthermore, AI has a reputation for making allegations only after carefully vetting them (Clark 2001, Cmiel 1999). The organization typically receives reports from its grass roots network and conducts a rigorous process to determine whether to go public with an allegation. If AI later learns that a published allegation was false, the organization publishes a retraction.[3]

3.2 Under-Reporting Bias

Under-reporting bias is well known to human rights researchers and is a problem for all data measuring either the allegation of human rights violations or the violations themselves

[3]This information is drawn from an interview with AI personnel conducted by Moore on 12 November 2008.

(Bollen 1986, 579–82; Spirer 1990; Cingranelli & Richards 2001; Goodman & Jinks 2003, 175–6; Hathaway & Ho 2004). We believe that this undercount is also relevant to existing large-N data on rights including the Political Terror Scale (Gibney & Dalton 1996) and the Cingranelli-Richards Human Rights Data (Cingranelli & Richards 2010). Because the PTS and CIRI generate ordered data rather than event data, however, the undercount is unlikely to be as much of a concern for researchers interested in drawing inferences about violations.

4 Reliability

Conrad et al. (2013b) describe the inter coder reliability tests conducted for the ITT project. Here we describe the selection and training of the ITT coders and then report two measures of reliability for each of the variables in the ITT SA data: the overall proportion of agreement measure (Fleiss 1971, 1981) and either Krippendorff's (2004) alpha (α_K) for variables with mutually exclusive values, or Light's (1971) kappa (κ_L) for variables whose values are non-mutually exclusive.

4.1 Selection and Training of ITT Coders

To receive an invitation undergraduate students needed to meet three criteria: excellent classroom performance, a grade point average of at least 3.7, and several Advanced Placement courses in US high school. Those who joined the project completed several weeks of individual and group coding training. Coders in training were required to pass a certification test (80% plus correct) before they were assigned cases to code. The certification test involved coding a small set of documents. It was graded by comparing the assigned values to a key that represented the "correct" coding of the documents. The key was produced by Conrad, and then reviewed by Moore (with discussion used to resolve any disagreement between the co-PIs). Once certified they earned US $12 per hour for coding.

We also periodically gave the coders who had certified what we called "certification checks." We designed these for the purpose of generating the data we used for calculating the inter coder reliability (ICR) of the coding scheme (see following subsection), but realized that we could also compare each coder's coding to a "correct" coding to ensure that each coder was continuing to code at an acceptable level (80% plus correct). In addition to using the data to conduct standard reliability scores, we evaluated each coder against a "correct" coding that was produced in the same manner we used to create the key for the certification test. Coders were told that if they failed a certification check they would be removed from coding until the studied and then re-certified (passed a new exam). No coders failed these certification checks.

Each individual coder was assigned a different country to code and was required to code all AI annual reports, press releases, and Action Alerts associated with their assigned country. AI also releases a number of topical reports that are not readily classified by country. Whenever a coder found such a "multi-country" document s/he would check to see whether that document had already been coded by another coder, and when it had not, s/he was assigned that document (and thereby coded it for all countries included in the document). Additional documentation related to the training and certification processes will be made available on the ITT Project website.

4.2 Reliability Testing and Scores

To assess reliability we emailed the trained coders documents composed of content taken directly from Amnesty International reports. To build a representative sample of content for our ICR tests we needed to ensure that they contained at least one allegation that would lead to assignment of a non-zero value to each of the many variables in the coding scheme. We also wanted to have regional and temporal variation. Thus, rather than give them whole documents, we selected portions of reports (some of which contained allegations, others of which did not) from different countries and years.

The coders were told that they were performing "certification checks" (the evaluations described above) to ensure that they maintained an adequate level of proficiency to continue coding the ITT data. Coders were instructed to perform content analysis on these documents in the same way as they would for a country they were coding, including entering information into a spreadsheet and documenting their coding notes. To collect the data for the inter-coder reliability (ICR) analysis we emailed the coders these documents twice during the Fall of 2009 and twice during the Spring of 2010. We analyzed the data in the standard way. That is, while certification checks involved comparison each individual's coding to a "correct" coding, the ICR analysis compared the coding of all of the coders with one another (and did not include the "correct" coding). Due to attrition, the Fall 2009 ICR analyses included 16 coders, while the Spring 2010 analyses included 15 coders and 14 coders respectively.

Table 1 provides information on the reliability of each of the variables in the ITT SA data. We report two measures of reliability for each of the variables in the ITT SA data: the overall proportion of agreement measure (Fleiss 1971, 1981) and either Krippendorff's (2004) alpha (α_K) for variables with mutually exclusive values, or Light's (1971) kappa (κ_L) for variables whose values are non-mutually exclusive. The equations for these three statistics, as well as the reasons we selected them, are described in Conrad et al. (2013b). Although there is no specific threshold to clear, a score greater than or equal to 0.8 is broadly considered sufficiently high to consider an instrument (and thus, the variable produced with it) reliable. All of the ITT SA variables except **Victim Type** have a (rounded) score greater than or equal to 0.8. We leave it to users to decide whether to use the SA Victim Type variable in their own analyses.

5 Who Gets Abused?

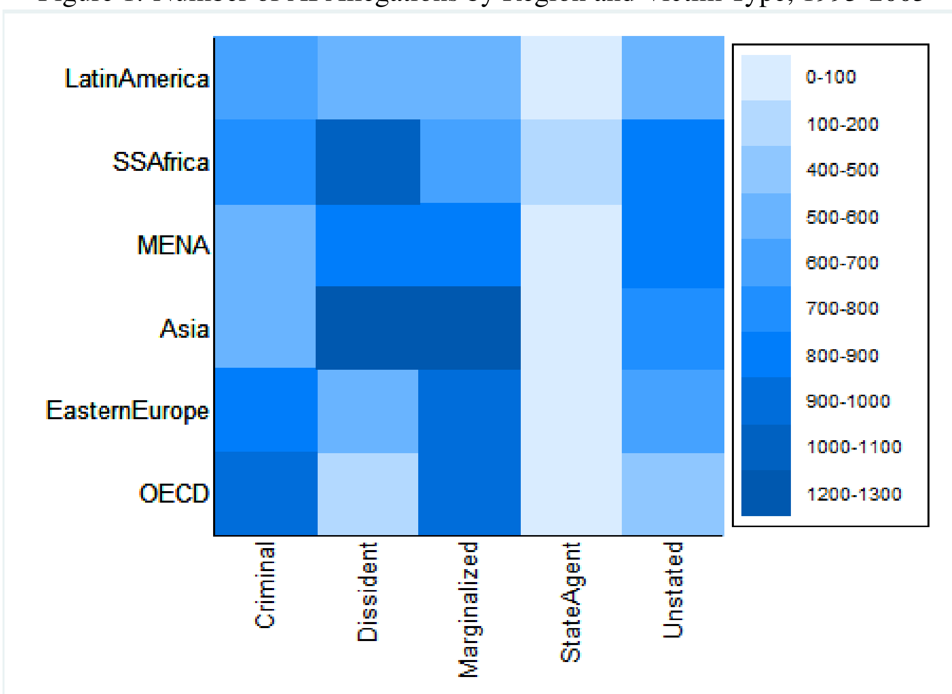
Members of Marginalized groups are the most frequently identified type of victim, being noted in over five thousand allegations. Dissident and Criminal victims are each mentioned in more than four thousand allegations, as are allegations in which AI fails to state a partic-

Table 1: **Reliability Scores for ITT Specific Allegation Variables**

	Proportion of Overall Agreement	α_K † / κ_L ‡
Order of Magnitude	0.951	0.953†
Victim Type	0.727	0.618‡
Expectation Torture Has/Will Occur	1.00	1.00†
Torture Type	0.919	0.932‡
Torture Death	0.901	0.846†
Agency of Control	0.893	0.882‡
Formal Compliant Filed	0.972	0.965†
Investigation of Torturers	0.950	0.942†
Outcome of Investigation	0.799	0.817‡
Location of Adjudication/Mediation	0.855	0.797†
Outcome of Adjudication/Mediation	0.838	0.803†
Transborder Torture	1.00	1.00†
Destination	1.00	1.00†

ular victim type. The regional distribution of victim types in AI allegations of torture and ill-treatment is depicted in Figure 1. Dissidents were identified as the victim in over one thousand allegations in both Asian and Sub-Saharan African countries and approximately another 900 in the Middle Eastern and North African countries. Criminals, on the other hand, were noted as the victim in more than 850 allegations in both the Western Europe and Eastern Europe and the Former Soviet Republic countries, with about 650 to 700 allegations in both Latin America and Sub Saharan Africa, and approximately 550 allegations against Asian and Middle Eastern and North African countries. AI failed to identify the type of victim most frequently in Sub-Saharan African and Middle Eastern and North African countries (∼850 times) and between 400 to 725 times in the other regions. Members of Marginalized groups were least frequently identified in Latin American and Sub-Saharan Africa (∼575 to 700 times), and most commonly in Asia (∼1,200 times).

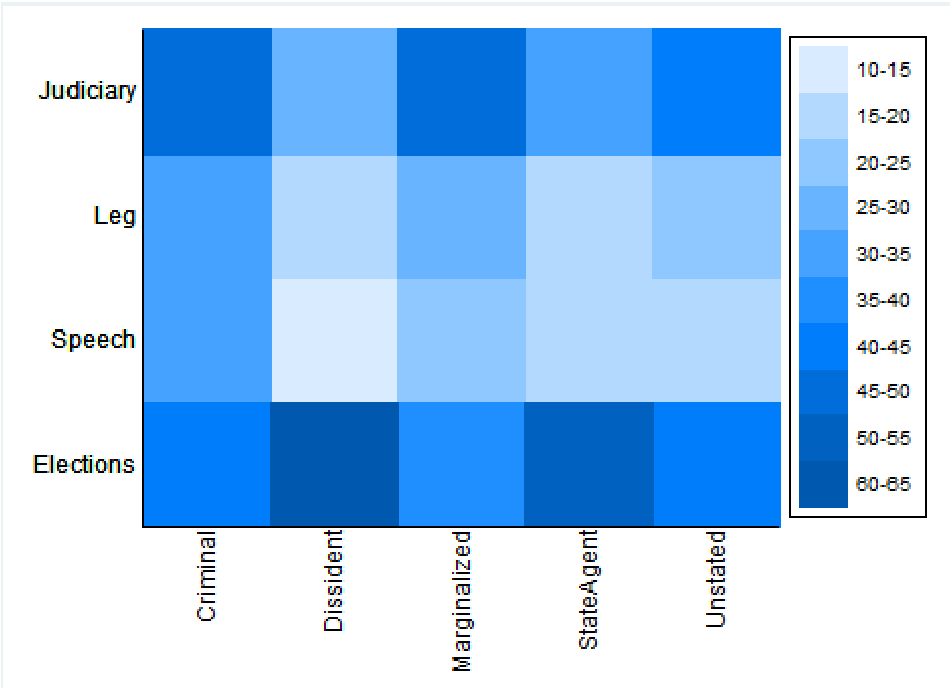
Figure 1: Number of AI Allegations by Region and Victim Type, 1995-2005



NOTES: Cell values describe the number of AI allegations against each victim type occurring in each geographic region.

Figure 2 reports the percentage of allegations against criminals, dissidents, marginalized individuals, state agents, or unstated victim types in states where there is a powerful judiciary, an elected legislature, free speech protections, or competitive elections, respectively. There are relatively few allegations against dissidents when these democratic institutions are present. Given that opposition in states with these institutions are more likely to mobilize, it is possible that AI makes fewer allegations concerning Dissident torture in such cases. The number of allegations involving different victim types do not vary strongly among states that hold elections: AI alleges about half of all allegations involving the torture of dissidents and state agents in countries where competitive elections are held and about half in countries where competitive elections are not held. Further, AI alleges about half of all allegations involving the torture of Criminals and Marginalized individuals occur in countries where a powerful judiciary is present. Free speech protections and an elected legislature exhibit strong relationships with fewer allegations of torture involving all types of victims, though particularly dissidents, state agents and unstated victims.

Figure 2: Prevalence of AI Allegations by Institution and Victim Type, 1995-2005



NOTES: Cell values describe the percentage of AI allegations of torture perpetrated against each victim type occurring in the presence of different political institutions.

6 Statistical Models of State Practice with ITT

In the article we observed that one can take a control variable approach to utilizing the ITT data to study states' (lack of) respect for the right to freedom from ill treatment and torture,[4] though there is a caveat. To see that this is so, let *AT* represent *unobservable* actual torture. Actual torture is unobservable due to a combination of the executive's incentive to hide any ill-treatment or torture that she (tacitly) approved, the agent of coercion's incentive to

[4]The alternative usage is to study AI's naming and shaming activity.

hide ill-treatment or torture that the executive did not (tacitly) approve, and the reporting agency's lack of shaming when a violation occurs. We wish to estimate the following regression, where X_i is a vector of variables implied by our hypotheses, and β is a vector of parameters to be estimated.

$$AT_i = \alpha_1 + X_i\beta + \varepsilon_i \tag{1}$$

If we let *OA* indicate observed torture allegations as reported in the ITT data, then actual torture is the sum of observed allegation and unobserved violations (UV): $AT \equiv OA + UV$. Existing work, such as Conrad & Moore (2010b), implicitly assumes that $UV = 0$. This is a very strong assumption that is unlikely to be true. More troubling, the ratio $OA : UV$ is unlikely to be constant across countries: groups like AI are likely to both observe different proportions of *AT* in different countries, and be more likely to report what they observe in some countries than others. The literature on the cross-pressuring incentives faced by INGOs (e.g., Berkovitch & Gordon 2008, Lake & Wong 2009, Hendrix & Wong 2013, Hill et al. 2013) suggests that OA is composed of two types: true allegations (TA) and false allegations (FA). We therefore revise as follows: $AT \equiv TA - FA + UV$. Unfortunately, *AT*, *TA*, *FA* and *UV* are all unobserved: all we observe is *OA*. What alternatives would be better than the implicit assumption that *FA* and *UV* equal zero?

Begin by letting $\upsilon_i = FA_i + UV_i + \eta_i$, and note that $AT_i = OA_i + \upsilon_i$, where $\upsilon_i \sim N(Z_i\theta, \sigma^2)$. Z_i is a vector of variables that describe the data generation process leading INGOs to fail to observe violations, choose not to report some that are observed, and issue allegations that turn out to be false. θ is a vector of parameters to be estimated. While it is common to see errors written with a mean of 0 and a variance of σ^2, the process that produces *FA* and *UV* is not randomly distributed with a zero mean. Instead, it can be modeled such that, conditional upon $Z_i\theta$, the residual is normally distributed with mean 0 and variance σ^2. To be complete, we write $\eta_i \sim N(0, \sigma^2)$. We substitute $(OA + Z_i\theta + \eta_i)$ for AT_i in equation 1, producing the regression:

$$(OA + Z_i\theta + \eta_i) = \alpha_1 + X_i\beta + \varepsilon_i \tag{2}$$

Subtracting $Z_i\theta$ and η_i from both sides of the equation produces a generic representation of the standard regression equation used in the literature,

$$OA = \alpha_1 + X_i\beta + Z_i\theta + (\varepsilon_i - \eta_i) \tag{3}$$

given the implicit assumption that θ and η_i are zero.[5] In such a case, Z is excluded from the regression. If we let $\xi_i = \varepsilon_i - \eta_i$, we get a standard generic representation of the regression we need to estimate $OA = \alpha_1 + X_i\beta + Z_i\theta + \xi_i$. This turns out to be a rather straight forward fix: to use the ITT data and draw inferences about the impact of covariates (X_i) on actual state torture, one simply must control for the covariates affecting the observation of allegations (Z_i) in empirical models predicting actual torture. Unfortunately, there is a drawback

[5] For ease of notation we present a generic linear case, but it extends to a Poisson or other event count model.

to this simple approach: any variable that is common to X_i and Z_i can only be entered into the regression once. As a result, the vector of parameter estimates is $\beta^* = \beta + \theta$. Unfortunately, there is no way to use the regression to assign specific values to β_i or θ_i (other than ad hoc identification assumptions). Substantive interpretation thus becomes murky for any variable included in both X_i and Z_i as one cannot assess the impact of such a variable on the states' performance v AI's likelihood of reporting. As such, researchers must be clear about that when interpreting results. In the following subsection we discuss how researchers can avoid that ambiguity by modeling both the data generating process for AI allegations and states' (lack of) respect for the CAT.

6.1 Undercount Models with Fully Identified Parameters

Cameron & Trivedi (1998, Section 10.5) describes a number of event count statistical models that have been developed to generate unbiased estimates of parameters when one is faced with an undercount. We briefly describe the modeling approach and intend to implement it, in a Bayesian setting, in the next iteration of our study. For ease of exposition, we develop the points in the context of a Poisson regression, but the point generalizes to the negative binomial model, which is a mixture of the Poisson and gamma distributions (Winkelmann 2008, pp. 134-38) and the Poisson-log normal model (Cameron & Trivedi 1998, pp. 128-38; Winkelmann 2008, pp. 132-34).[6]

A brief aside on zero-inflated models (e.g., Long 1997, pp. 142-47) might prove useful. These models permit one to account for biased undercounts of a specific value: 0. They assume, however, that all other observed values (i.e., 1 - ∞) are unbiased. As such, they are not useful for the general case of an undercount. Yet, "a zero-inflated count model is a special case of a finite mixture model" (Cameron & Trivedi 1998, pp. 128). We are interested in the finite mixture models that can be used to model undercounts.[7]

One can represent the Poisson regression model as

$$Pr[Y = y] = \frac{e^{-\mu}(\mu)^y}{y!}, y = 0, 1, 2, \ldots \tag{4}$$

where y is an observed number of events over a given unit of time and μ is the mean of y. Note that Y is the true number of events while y is the observed number of events that are recorded in our dataset. Using the notation from our study, $Y = AT$ and $y = TA$.

[6]The Poisson log-normal has no closed form solution (Cameron & Trivedi 1998, pp. 143), and this has limited its use in estimation. However, unlike frequentist methods, Bayesian (and other simulation) estimation approaches do not require a closed form solution, and Winkelmann (2008, p. 134) reports that because "it fits the data often much better than the negative binomial model... the previous neglect of the Poisson-log-normal model in the literature should be reconsidered in future applied work."

[7]Finite mixture models have not yet become widely used in political science, in part because their use by and large requires the researcher to program statistical software to do the estimation. Deb Partha's FMM module for Stata (`http://econpapers.repec.org/software/bocbocode/s456895.htm`) makes it easier to estimate regression models that mix a large variety of distributions, and given the popularity of Stata in political science these models may be more widely adopted.

Recall that homogeneous negative bias in a dependent variable shrinks the size of estimated parameters, but does not otherwise negatively impact estimates (e.g., Achen 1986, chap. 4). Our problem is that we have heterogenous bias: we cannot reasonably assume that our undercount is either uniformly or normally distributed. We do not know what distribution it has, but we can be confident it is neither of those. Further, we have theory to help us make a case for what variables impact the extent to which events become allegations.

Cameron & Trivedi (1998, p. 313) explain that "the basic idea [is] that modeling the recording process may result in improved inference about parameters of interest." To begin Cameron and Trivedi introduce a new parameter, π, which represents the probability that an event which occurs is observed and recorded.[8]

$$Pr[Y=y] = \frac{e^{-\mu\pi}(\mu\pi)^y}{y!}, y = 0, 1, 2, \ldots \tag{5}$$

When $\pi = 1$ equation 5 reduces to equation 4, and $y = Y$. Our problem is that we are certain that $y < Y$, which is to say: $\pi < 1$. If we were able to argue that π was either uniformly distributed (i.e., every event in all country-years had the same probability of being observed and recorded) or normally distributed around 0.5 (i.e., every event in all country-years had, on average, a 50% chance of being observed and recorded, and was equally likely to have chance greater than or less than 50%), then we would be in the well known situation where we would be generating downward biased estimates by assuming that $\pi = 1$ and estimating a regression based on equation 4.

In our study we cannot reasonably assume π is homogeneous across countries: INGOs like AI are not equally likely to observe and publish all violations of the CAT that occur in the sundry government detention centers throughout the world. That is, we cannot reasonably assume that the value of π for an event in Argentina in year t is equal (on average) to the value of π for an event in Norway in year t, which is also equal (on average) to the value of π for an event in North Korea in year t, and so on. We do, however, have theory about how INGOs like AI produce allegations that permit us to identify covariates that will impact the value of π (e.g., Hill et al. 2013). That is the key insight: we are able to introduce a parameter that impacts the chance that AI produces an allegation, and then estimate its value as a function of covariates. Doing so will allow us to disentangle the effect of covariates on both torture violations and their allegations.[9] Further, we need not assume one single value for π, but can instead estimate country-specific values of π, much as one does in fixed and random effects models.

We leave to Cameron & Trivedi (1998, section 10.5) the details on the generalization of these models to the negative binomial case, as well as discussion of whether errors across the equations are correlated (we will need to assume that they are).[10] We plan to estimate

[8] Cameron & Trivedi (1998, pp. 313) refer to models that introduce π as binomial thinning process models (these were initially introduced in a time-series context; see pp. 234-36).

[9] Although we assume that institutions like elections only affect torture *violations* above, these econometric models will allow us to determine whether or not that is the case.

[10] We also plan to account for correlated errors across stealth, scarring, and unstated torture types.

these models for the next version of our study.

7 Additional Figures

The figures below provide additional descriptive information about the ITT SA data.

Figure 3: Number of AI Allegations by Region, 1995-2005

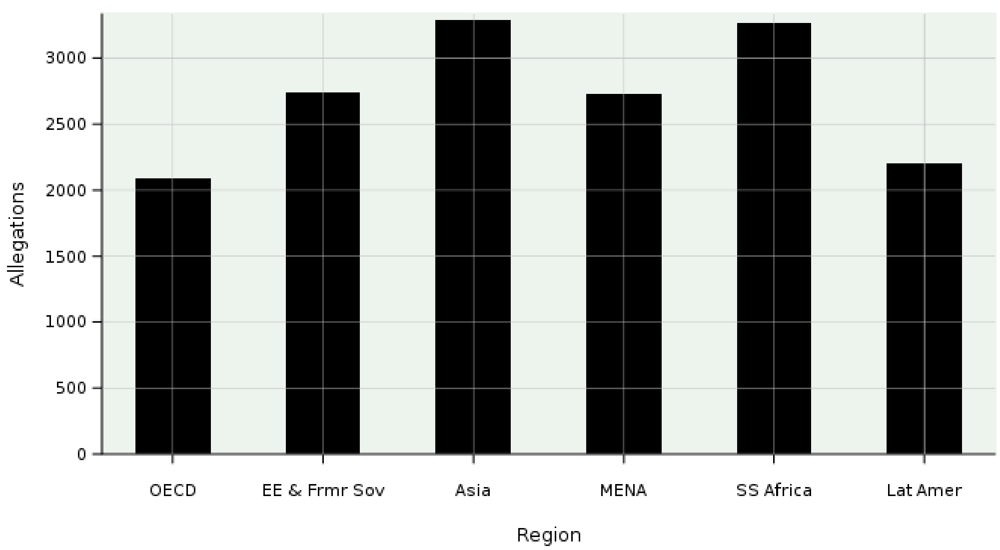
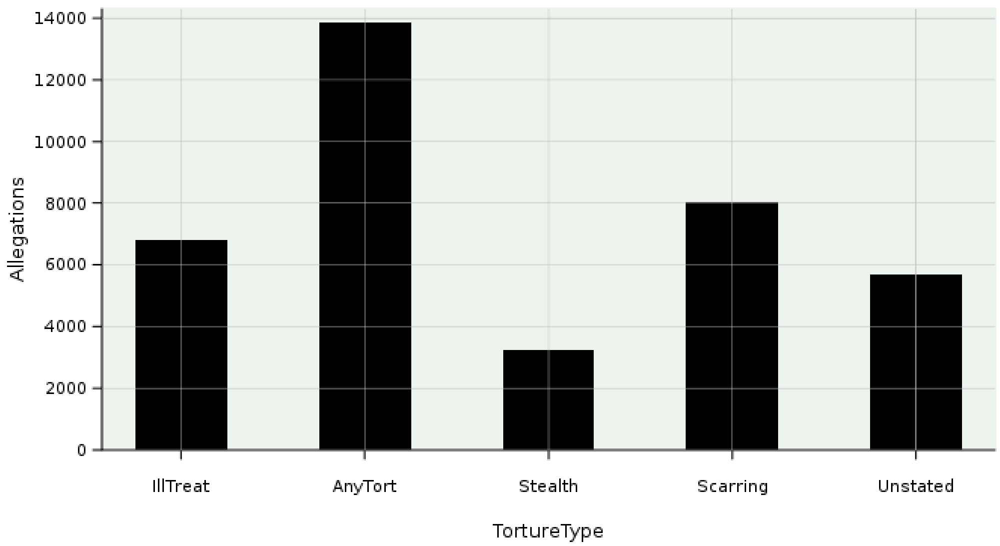
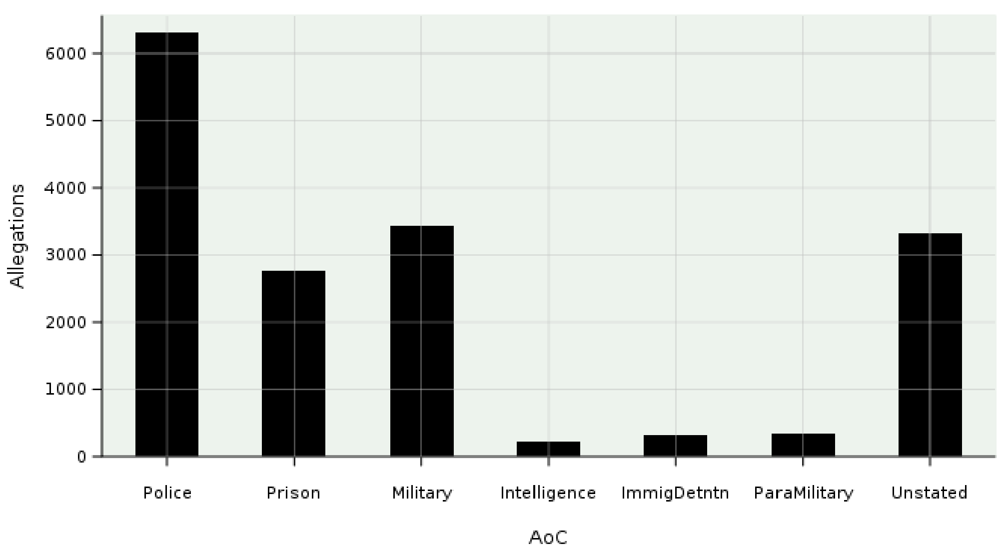


Figure 4: Number of AI Allegations by Torture Type, 1995-2005



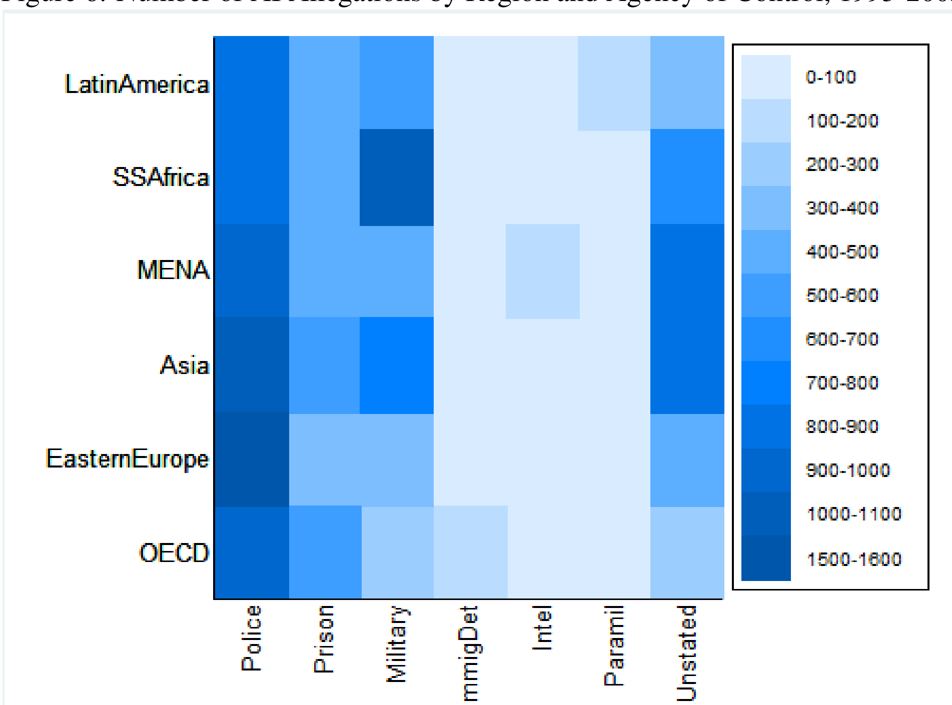
NOTES: Torture types are not mutually exclusive. Any Torture includes allegations of at least one torture type.

Figure 5: Number of AI Allegations by Government Agency of Control, 1995-2005



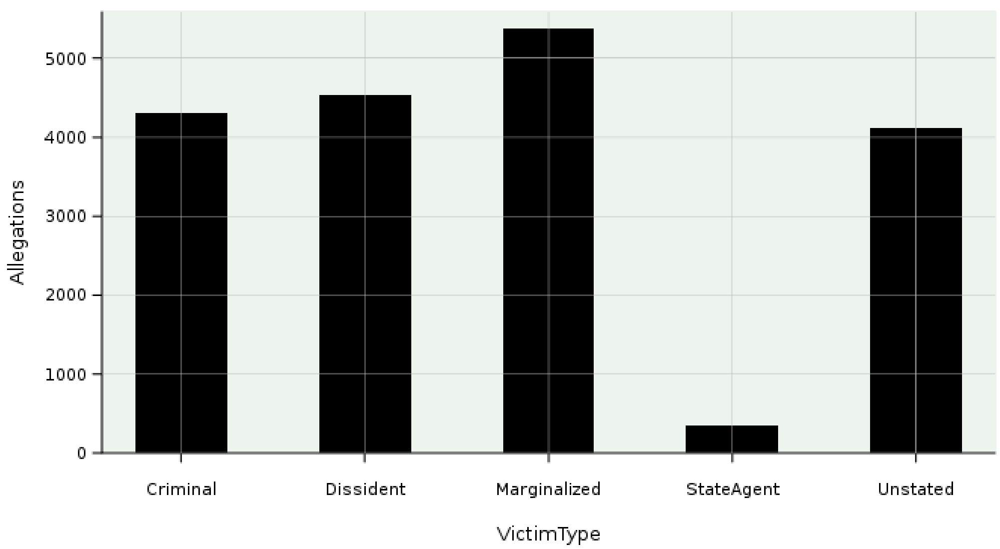
NOTES: Agencies of Control are not mutually exclusive.

Figure 6: Number of AI Allegations by Region and Agency of Control, 1995-2005



NOTES: Cell values describe the number of AI allegations of each torture type by alleged Agency of Control.

Figure 7: Number of AI Allegations by Victim Type, 1995-2005



NOTES: The types are not mutually exclusive: a given (group of) detainee(s) may belong to more than one type of group.

References

Achen, Christopher H (1986) *The Statistical Analysis of Quasi-experiments*. Berkeley University of California Press.

Berkovitch, Nitza & Neve Gordon (2008) The political economy of transnational regimes: The case of human rights. *International Studies Quarterly* 52(4): 881–904.

Bollen, Kenneth A (1986) Political rights and political liberties in nations: An evaluation of human rights measures, 1950 to 1984. *Human Rights Quarterly* 8(4): 567–591.

Cameron, A. C & Pravin K Trivedi (1998) *Regression Analysis of Count Data*. New York Cambridge University Press.

Cingranelli, David L & David L Richards (2001) Measuring the impact of human rights organizations. In: C.W. Welch (ed.) *NGOs and Human Rights: Promise and Performance*. Philadelphia University of Pennsylvania Press , 225–237.

Cingranelli, David L & David L Richards (2010). The Cingranelli-Richards (CIRI) Human Rights Dataset. http://www.humanrightsdata.org.

Clark, Ann M (2001) *Diplomacy of Conscience: Amnesty International and Changing Human Rights Norms*. Princeton Princeton University Press.

Cmiel, Kenneth (1999) The emergence of human rights politics in the united states. *The Journal of American History* 86(3): 1231–1250.

Conrad, Courtenay R; Jillienne Haglund & Will H Moore (2013a) Disaggregating torture allegations: Introducing the ill-treatment and torture (itt) country-year data. *International Studies Perspectives* 14(13): 199 – 220.

Conrad, Courtenay R; Jillienne Haglund & Will H Moore (2013b). The Ill-Treatment & Torture (ITT) Data Project Intercoder Reliability Analysis. Merced and Tallahassee: Ill Treatment and Torture Data Project.

Conrad, Courtenay R & Will H Moore (2010a). The Ill-Treatment & Torture (ITT) Data Project Coding Rules & Norms. Merced and Tallahassee: Ill Treatment and Torture Data Project.

Conrad, Courtenay R & Will H Moore (2010b) What stops the torture? *American Journal of Political Science* 54(2): 459 – 476.

Conrad, Courtenay R & Will H Moore (2011). The Ill-Treatment & Torture (ITT) Data Project Specific Allegations Data User's Guide. Merced and Tallahassee: Ill Treatment and Torture Data Project.

Fleiss, Joseph L (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378–382.

Fleiss, Joseph L (1981) The measurement of interrater agreement. In: J.L. Fleiss; B. Levin & M.C. Paik (eds.) *Statistical methods for rates and proportions*. New York Wiley , 212–236.

Gibney, Mark & Matthew Dalton (1996) The political terror scale. *Policy Studies and Developing Nations 4*: 73–84.

Goodman, Ryan & Derek Jinks (2003) Measuring the effects of human rights treaties. *European Journal of International Law* 14(1): 171–184.

Hathaway, Oona A & Daniel E Ho (2004) Characterizing measurement error in human rights. Paper presented at the annual meeting of the American Political Science Association.

Hendrix, Cullen S & Wendy H Wong (2013) When is the pen truly mighty? regime type and the efficacy of naming and shaming in curbing human rights abuses. *British Journal of Political Science* 43(3): 651–672.

Hill, Daniel W; Will H Moore & Bumba Mukherjee (2013) Information politics v organizational incentives: When are ingo's "naming and shaming" reports biased? *International Studies Quarterly* 57(2): 219–232.

Krippendorff, Klaus (2004) *Content analysis: An introduction to its methodology*. Thousands Oaks Sage Publications.

Lake, David A & Wendy H Wong (2009) The politics of networks: Interests, power, and human rights norms. In: Miles Kahler (ed.) *Networked Politics: Agency, Power, and Governance*. Ithaca Cornell University Press , 127–150.

Light, Richard J (1971) Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin* 76(5): 365–377.

Long, J. S (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks Sage Publications.

Rejali, Darius (2007) *Torture and Democracy*. Princeton, New Jersey Princeton University Press.

Spirer, Herbert F (1990) Violations of human rights–how many? *American Journal of Economics and Sociology* 49(2): 199–210.

Winkelmann, R. (2008) *Econometric analysis of count data* Springer Verlag.