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

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. Following the training, coders were required to pass a certification test before they were permitted to code ITT data. Once certified they earned US $12 per hour for coding. Four certification checks were administered during the project.

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.

2 Reliability

We conducted a series of intercoder reliability (ICR) checks throughout the coding process. Every coder in the project was told that to maintain their status with the project they had to pass what we called "certification checks."[1] These certification checks were emailed

[1] Coders in training were required to pass a certification test (80% plus correct) before they were assigned cases to code. We also periodically gave the coders who had certified what we called "certification checks." These were actually inter coder reliability tests, though we did not want to refer to them as such. 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 a certification check.

to trained coders and contained content taken directly from Amnesty International reports with the intention to assess the reliability of specific variables in the ITT data. Coders were told that they were performing "certification checks" or evaluations 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.[2] These ICR checks were performed twice in the Fall of 2009 and twice in the Spring of 2010.[3] The text from the four certification checks are below.

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, Haglund and Moore (2013).

3 Undercount Models with Fully Identified Parameters

Cameron and 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

[2]The content of the certification checks is included in the appendix to this document.
[3]The Fall 2009 ICR analyses included 16 coders, while the Spring 2010 analyses included 15 coders and 14 coders respectively.

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†

(Winkelmann 2008, pp. 134-38) and the Poisson-log normal model (Cameron and Trivedi 1998, pp. 128-38; Winkelmann 2008, pp. 132-34).[4]

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 and Trivedi 1998, pp. 128). We are interested in the finite mixture models that can be used to model undercounts.[5]

One can represent the Poisson regression model as

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

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 $y = TA$. Recall that homogeneous negative bias in a dependent variable shrinks the size of estimated parameters, but does not otherwise negatively impact estimates (e.g., **?**, 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 that occur will become allegations.

[4]The Poisson log-normal has no closed form solution (Cameron and 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."

[5]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.

Cameron and 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.[6]

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

When $\pi = 1$ equation 2 reduces to equation 1, 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 1.

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, Moore and Mukherjee 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

[6]Cameron and 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).

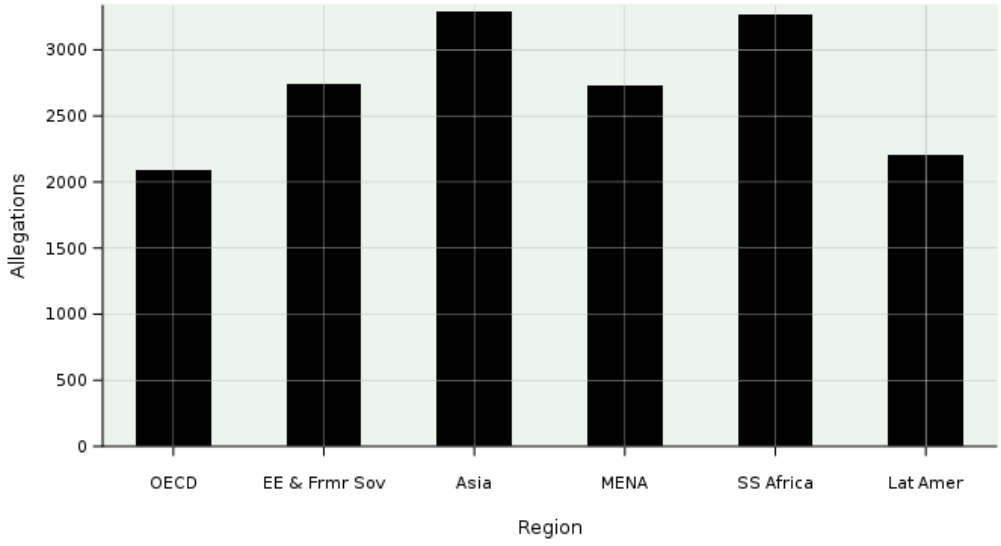
the effect of covariates on both torture violations and their allegations.[7] Further, we need not assume one single value for π, but can instead estimate country-specific values of π, much as one does in fixed and random effects models.

We leave to Cameron and 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).[8] We plan to estimate these models for the next version of our study.

4 Additional Figures

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

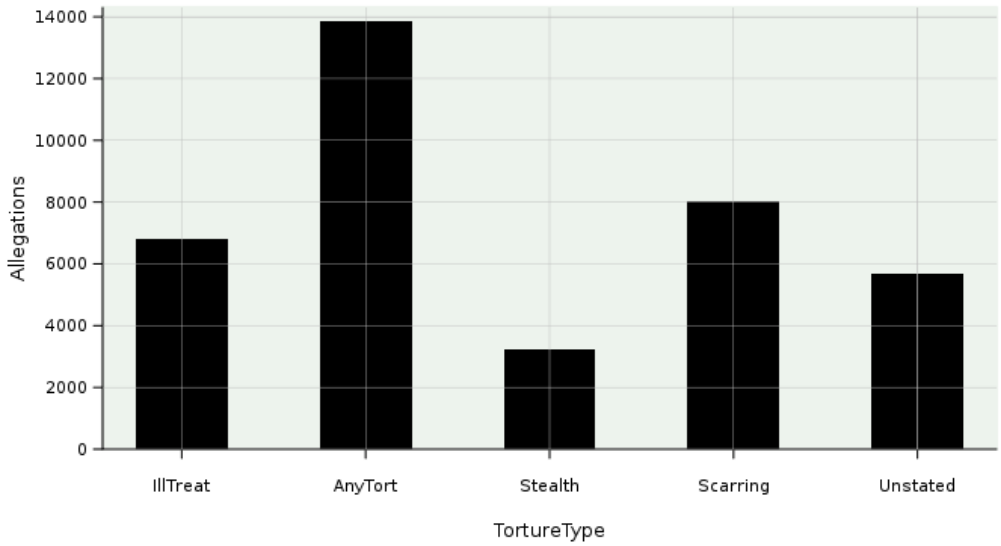
Figure 1: Number of AI Allegations by Region, 1995-2005



[7] 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.

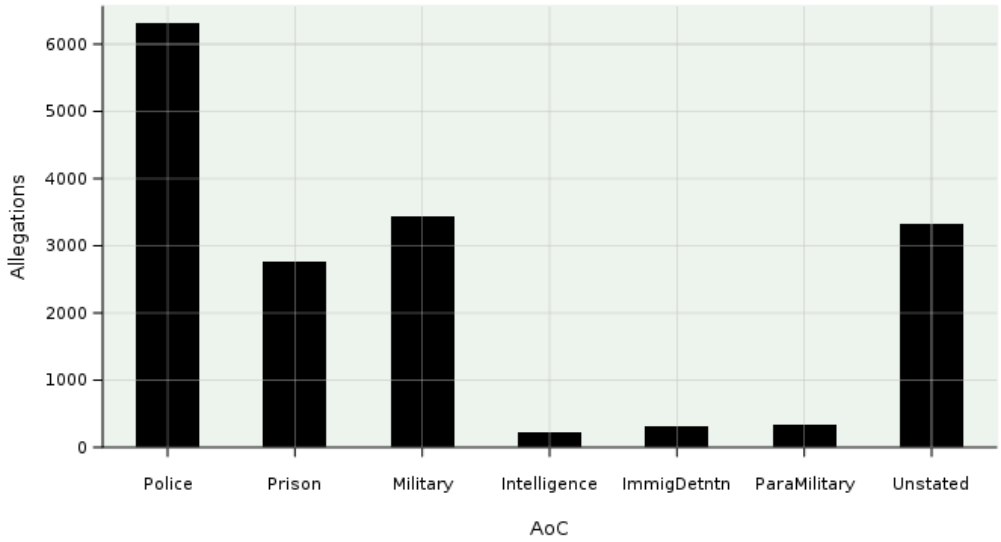
[8] We also plan to account for correlated errors across stealth, scarring, and unstated torture types.

Figure 2: Number of AI Allegations by Torture Type, 1995-2005



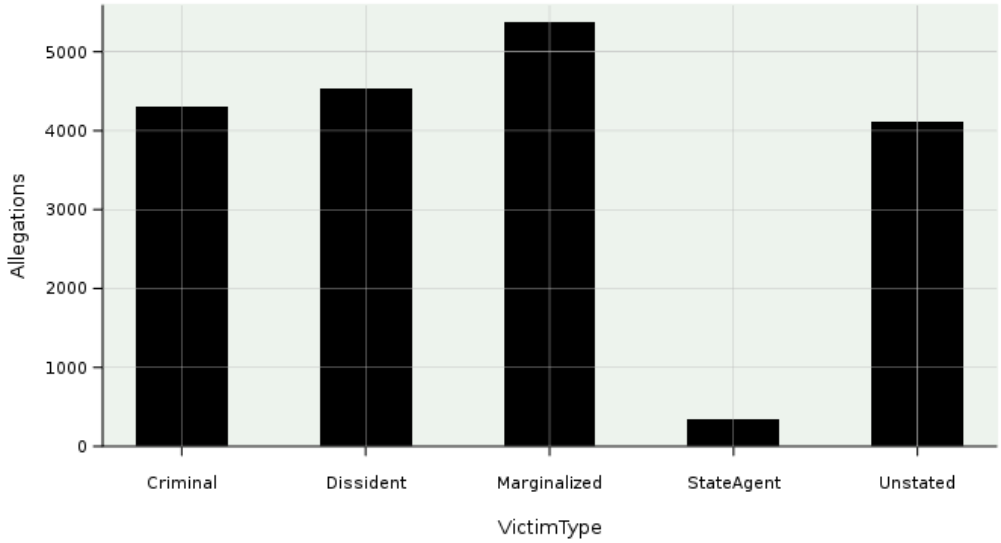
NOTES: Torture types are not mutually exclusive. Any Torture includes allegations of at least one torture type.

Figure 3: Number of AI Allegations by Government Agency of Control, 1995-2005



NOTES: Agencies of Control are not mutually exclusive.

Figure 4: 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

Cameron, A.C. and PK Trivedi. 1998. *Regression analysis of count data*. Cambridge Univ Press.

Conrad, Courtenay R., Jillienne Haglund and Will H. Moore. 2013. *The Ill-Treatment & Torture (ITT) Data Project Intercoder Reliability Analysis*. Merced and Tallahassee: Ill Treatment and Torture Data Project.
URL: *http://www.politicalscience.uncc.edu/cconra16/UNCC/Under_the_Hood.html*

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 *Statistical methods for rates and proportions*, ed. J.L. Fleiss, B. Levin and M.C. Paik. New York: Wiley pp. 212–236.

Hill, Daniel W., Will H. Moore and 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.

Light, Richard J. 1971. "Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives." *Psychological Bulletin* 76(5):365–377.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage Publications.

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