Supporting Information for: "Torture and the Limits of Democratic Institutions"

October 7, 2016

Contents

1 Motivation for Undercount Model

The ITT data include the population of AI *allegations* of torture made against a given state from 1995 to 2005 (Conrad et al., 2014). But our theory is not about allegations of human rights violations; we wish to draw inferences about *actual* state torture. Further, the allegations making up this population are not randomly drawn from the underlying distribution of actual instances of torture. They are an undercount of actual torture violations.[1] To provide clarity we define some terms.

[1] 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 (Bollen, 1986, 579–82; Spirer, 1990; Cingranelli & Richards, 2001; Goodman & Jinks, 2003, 175–6; Clark & Sikkink, 2013).

Let AT represent the number of cases of actual torture. Some portion of these cases are 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 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.

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

Let OA indicate observed torture allegations as reported in the ITT data. Actual torture is the sum of observed allegations 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 more of AT in some countries and more likely to report their findings in some countries than others. Unfortunately, UV is unobserved: all we observe is OA. The question, then, is whether there are alternatives superior to the implicit assumption that UV is equal to zero?

Begin by letting $\upsilon_i = UV_i + \eta_i$, 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 affect whether INGOs fail to observe violations, or observed violations and issue an allegation. 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 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:

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

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,

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

given the implicit assumption that θ and η_i are zero.[2] In such a case, Z is excluded from the regression. If we let $\xi_i = \epsilon_i - \eta_i$, 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, we might control for the covariates affecting the observation of allegations (Z_i) in our empirical models predicting actual torture (see Conrad et al., 2013, 2014). Because our dependent variables are counts of torture allegations, we could turn to a negative binomial model and include the relevant covariates in our specification.[3]

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

[3]Ordinary least squares (OLS) regression produces biased coefficient estimates when the dependent variable is categorical (Long & Freese, 2001). The negative binomial model adds a parameter to account for dispersion and reduces to the Poisson when this parameter is equal to zero (Long & Freese, 2001).

Two issues make this "OLS with control variables" approach unappealing. First, note that 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, and researchers must exercise care when interpreting those results. Second, this approach is not the only option when estimating a count variable (see Hill et al., 2013).

Cameron & Trivedi (1998, chapter 10) discuss a large number of event count regression models that permit one to model measurement error. Section 10.5 specifically addresses regression models one can estimate to model undercounts.[4] One of the chief advantages of these models is that they make it possible to identify the impact of variables researchers believe influence both the underlying data generation process and the likelihood that we observe the event.[5]

1.1 Undercount Models with Fully Identified Parameters

Cameron & Trivedi (1998, Section 10.5) describe a number of event count statistical models that have been developed to generate unbiased estimates of parameters when one is faced with an undercount. 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

[4]These models have not, to our knowledge, migrated to routines in large statistical software packages, and that likely explains their limited usage in the social sciences.

[5]Zero-inflated models are a special case of these models. Note that they only permit one to model zeros that should have a non-zero value, not other values that are under-counted.

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).

The p.m.f. of the Poisson distribution can be written as:

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

where y is the number of events that occur over a given unit of time and μ is the mean number of events. 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 $y = OA$. 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 the undercounting process is uniform across observations. Further, we have theory to help us make a case for what variables impact the extent to which events that occur will 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.[6]

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

[6]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).

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 constant across the country-years in our sample (i.e., the probability of observing an event if one occurs is the same in all country-years), then we would be in the well known situation of 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 to the value of π for an event in Norway in year t, which is also equal to the value of π for an event in North Korea in year t, and so on. If this assumption is not met then parameters from a standard count model are likely to be biased downward, and the bias will be larger as the ability of AI to acquire information decreases (see Feinstein, 1990, p. 247). 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 represents the chance that AI produces an allegation if torture occurs, and then estimate its value as a function of covariates. This means we need not assume one single value for π, but can instead estimate country-year-specific values of π. Doing so allows us to disentangle the effect of covariates on both torture violations and their allegations.[7] We refer the reader to Cameron & Trivedi (1998, section 10.5) for the details

[7] Although we assume that institutions like elections only affect torture *violations* above, these econometric

on the generalization of these models to the negative binomial case. The likelihood function for the undercount negative binomial model is presented in the main text for this study.

2 Descriptive Statistics

Here we present more detailed descriptive statistics for our dependent variables and our key independent variables. Figure 1 displays descriptive statistics for scarring torture allegations, ditinguishing between countries that hold competitive elections and those that do not. For each country, the median value for scarring torture allegations for all years in the sample (for which that country falls into the "elections" or "no elections" category) is shown as a dot, with the inter-quartile range shown as a line. Notice that there appears to be little difference in the number of allegations across the two groups, except near the top of the graph. Countries that do and do not hold free and fair elections often have precisely the same values for much of their interquartile range: the lower quartile is one, the upper is four. As Figure 1 suggests, the largest differences between the two groups emerge only in the upper range of the distributions, with "no election" countries producing nine allegations of scarring techniques in the 75th percentile and a maximum of 65, while "election" countries yield 10 at the 75th percentile and a maximum of 130 allegations.

As noted, democracies were the first to develop methods of clean or clean torture that leave few (or no) marks on the body of the victim (Rejali, 2007, 69-78; see also Ron, 1997). Unlike scarring torture, which leaves lesions and/or scars on the victim's body,[8] clean

models will allow us to determine whether or not that is the case.

[8]Scarring torture includes (but is not limited to) burning, beating, cutting, whipping, boiling, sexual

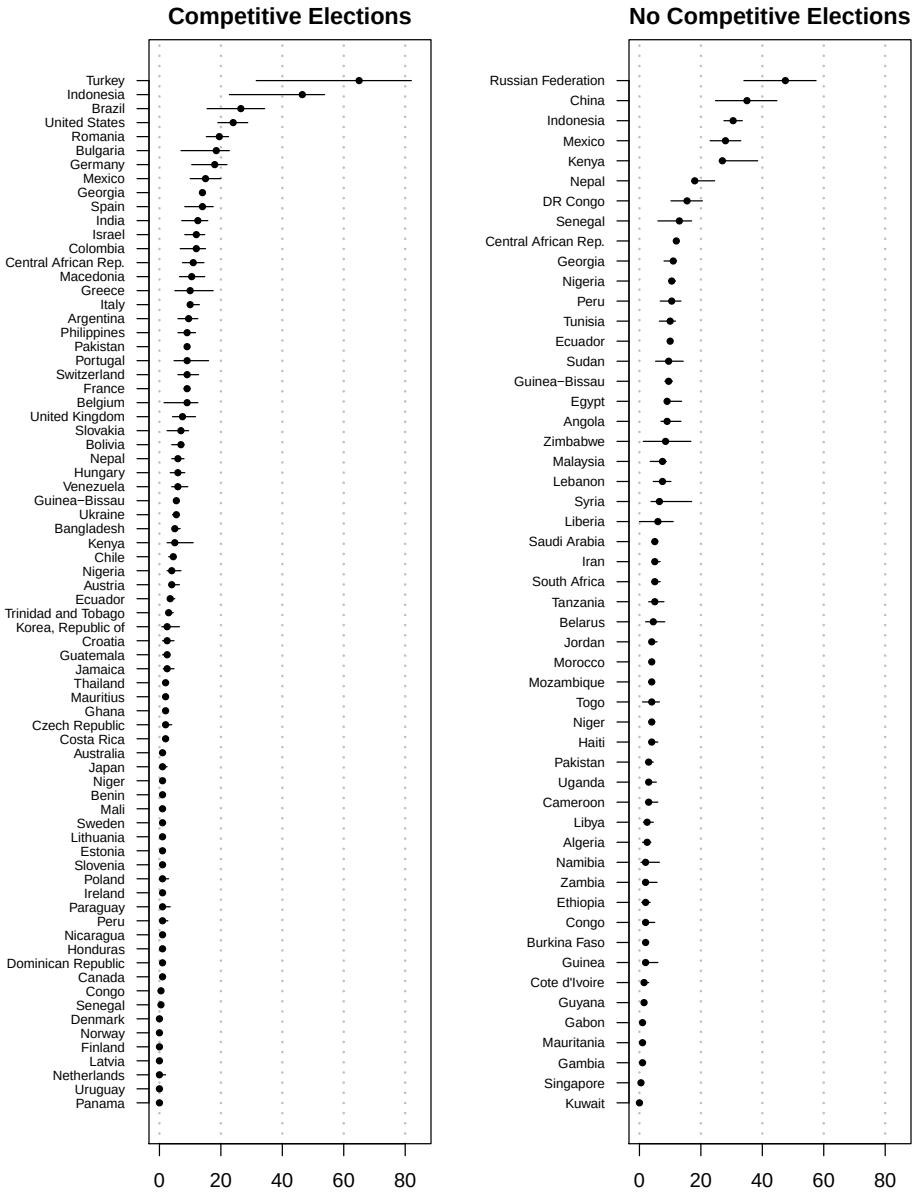

Figure 1: **Scarring Torture Allegations by Elections**. *Country-years which have a value of 1 for the Cheibub, Gandhi, and Vreeland measure are show on the left, country-years with a value of 0 are shown on the right. The median value for each country across all years in the sample is shown as a dot, with the inter-quartile range shown as a line.*

torture is purposefully and carefully executed so as not to leave visible marks on the victim.[9] Descriptive statistics for clean torture allegations in countries with and without elections are shown in Figure 2. The ITT Specific Allegations data indicate that AI has called out countries that hold elections for clean techniques less frequently than it has those that do not, with 50th and 75th percentile values of one and three for the former, and two and five for the latter.

We also argue in the paper that powerful courts[10] will constrain governments' use of scarring torture, and stimulate use of clean torture. Put plainly, we believe that powerful courts, not elections, are the primary vehicle that stimulates the adoption of clean techniques. The argument is wholly consistent with Rejali's (2007) case studies of the development of clean torture in the US, France and UK as well as Ron's (1997) account of the Israeli case. Yet, what does the ITT data suggest?

Treating Linzer and Staton's (2012) measure of judicial power as a binary variable,[11] we divide the sample into low/high judicial power country-years, and display descriptive statistics for scarring torture allegations for each group in Figure 3. Countries with powerful courts have the same median and third quintile as those without (four and 10, respecitvely),

abuse (to include rape), abuse using animals (e.g., allowing dog bites), maiming, and disfiguring (Conrad & Moore, 2010a).

[9]Clean/clean torture includes (but is not limited to) electrotorture, beating with instruments, beating on body parts so as not to leave marks, water torture, dry choking, climatized air, exhaustion exercises, positional torture and devices, restraints, irritants, sleep deprivation, noise, sensory deprivation, purposefully withholding food/water/medication, isolation from all human beings, forced feeding (Conrad & Moore, 2010a).

[10]Staton (2010) defines the power of a court over two dimensions: its independence (i.e., ability to make rulings based on the preferences of the justices rather than those of stake holders outstide of the court) and effectiveness (i.e., the extent to which other actors comply with court rulings). We adopt that definition here (see, also, Staton & Moore, 2011).

[11]We created a cutoff of 0.47, which is the median. Their variable is continuous between 0 and 1, with higher values representing greater power.

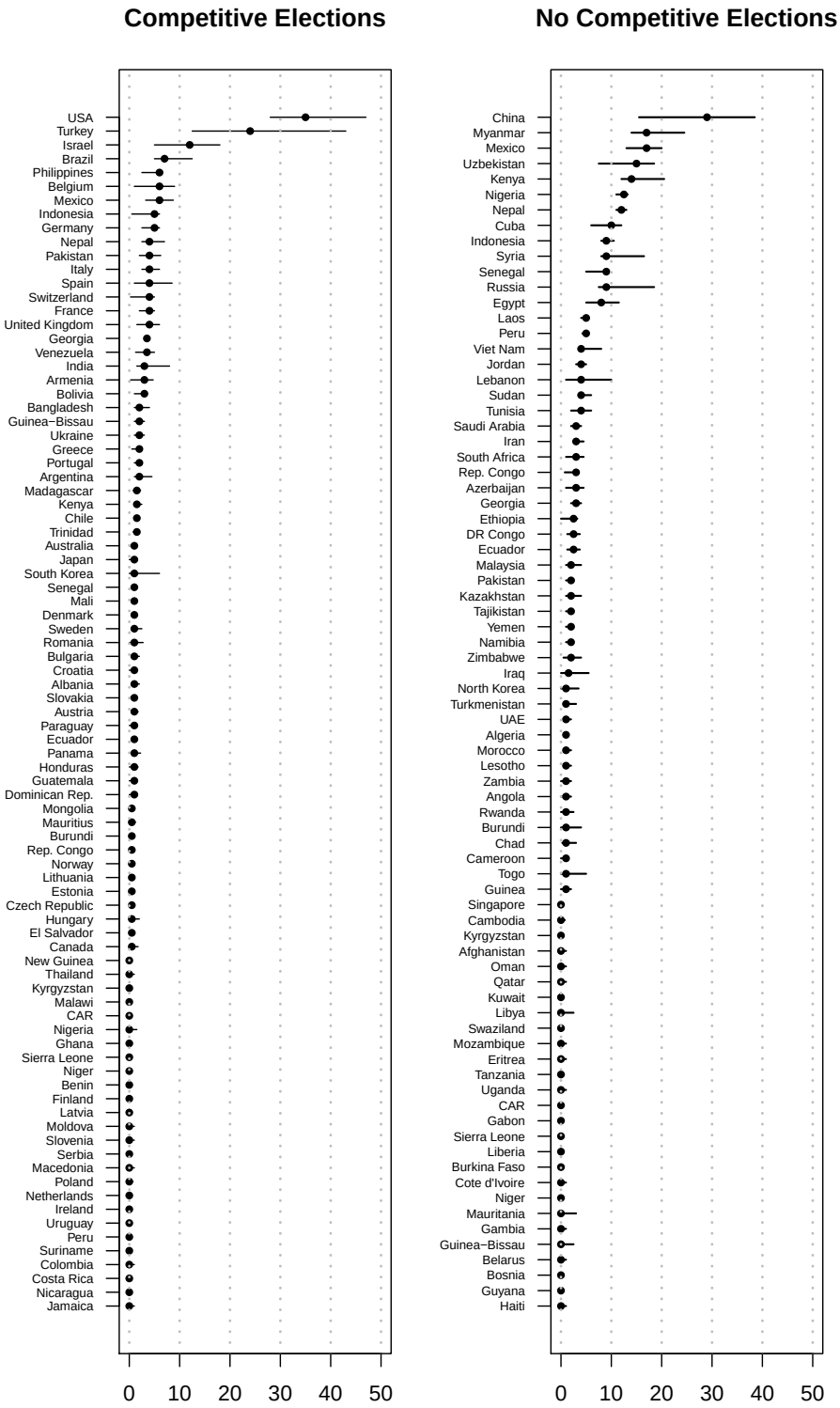

Figure 2: **Clean Torture Allegations by Elections**. *Country-Years which have a value of 1 for the Cheibub, Gandhi, and Vreeland measure are show on the left, country-years with a value of 0 are shown on the right. The median value for each country across all years in the sample is shown as a dot, with the inter-quartile range shown as a line.*

but those without have a higher maximum (130 v 75). We find the same pattern when we turn to clean torture: 50th percentiles of one and 75th of four for both groups, with a maximum of 80 allegations for states with powerful courts and 58 for those without.

Unlike the impression produced by a bivariate examination of the association between elections and torture techniques, the bivariate examination of judicial power and torture techniques cuts against our argument. That is, countries with powerful judiciaries are no more likely to be called out by AI for either scarring or clean techniques than are those with weak ones, except at the high end of the range, and there it is the countries with powerful courts that generate the most allegations of each type of technique.

3 ITT Data, Undercount Bias and the Moving Standards of INGOs

Figure 5 displays coefficient estimates from the detection equation. These results provide information about how variables in the model affect the probability that AI would have enough information to issue an allegation of torture if torture occurred. For example, the results for INGOs suggest that, the larger the number of human rights INGOs that have offices in a given country, the higher is the probability that, if torture occurs, AI will have enough information to produce an allegation of torture; the coefficients are both positive, and the confidence intervals do not contain zero. Perhaps surprisingly, freedom of speech and the press has no discernible impact on AI's ability to detect cases of clean or scarring torture.

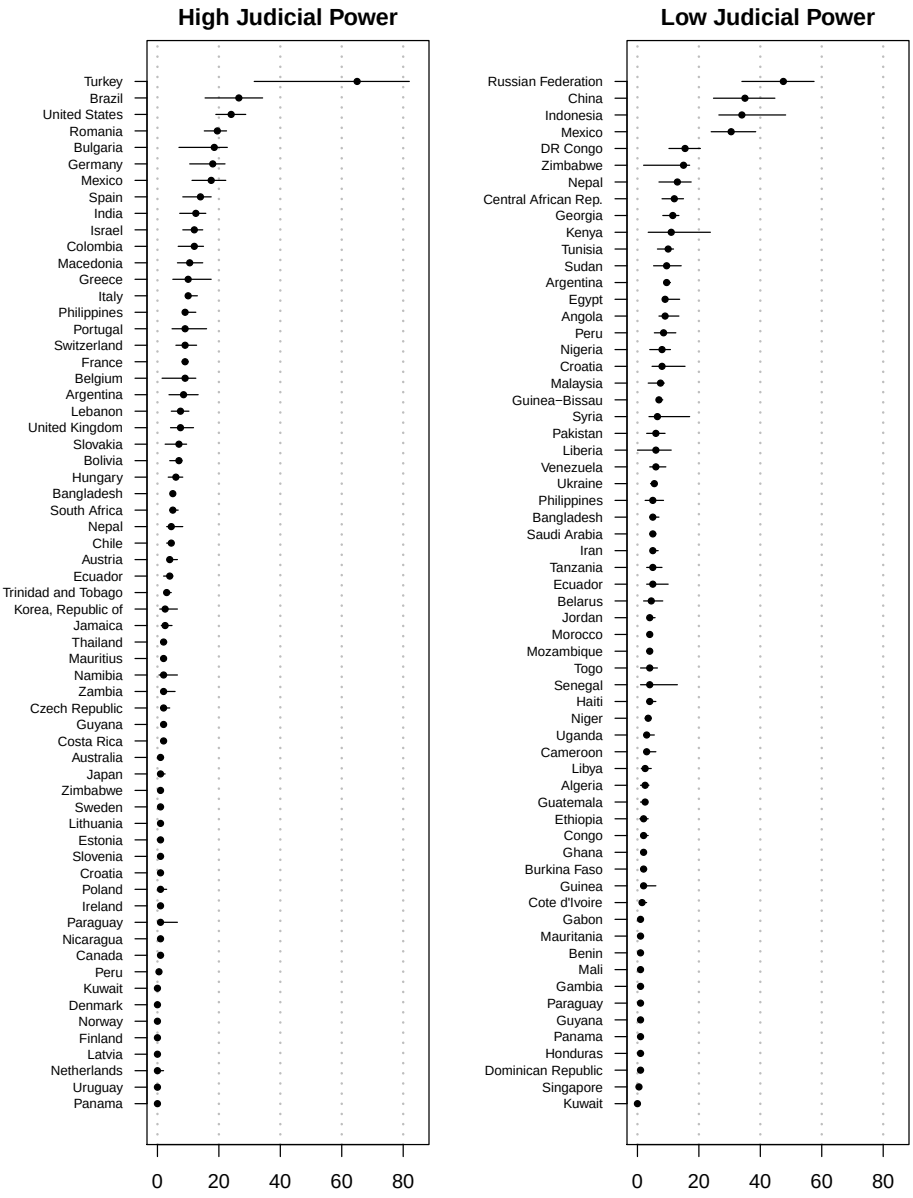

Figure 3: **Scarring Torture Allegations by Judicial Power**. *Allegations in Country-Years which have a value* $>= 0.47$ *for the Linzer-Staton measure of judicial power are shown on the left, country-years with values* < 0.47 *are shown on the right. The median value for each country across all years in the sample is shown as a dot, with the inter-quartile range shown as a line.*

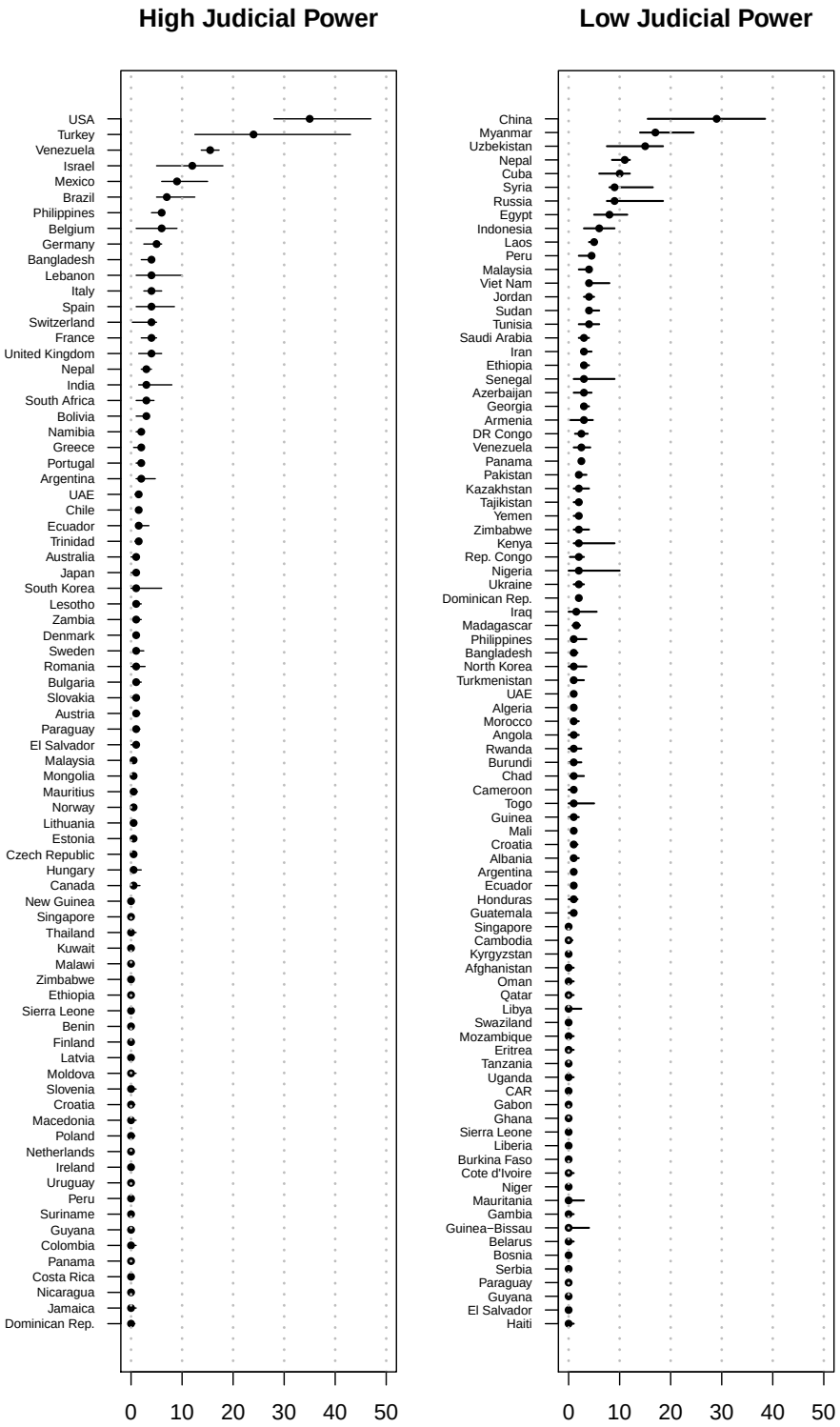

Figure 4: **Clean Torture Allegations by Judicial Power**. *Country-Years which have a value >= 0.47 for the Linzer-Staton measure of judicial power are shown on the left, country-years with values < 0.47 are shown on the right. The median value for each country across all years in the sample is shown as a dot, with the inter-quartile range shown as a line.*

We suspect that is due to the inclusion of the lagged human rights score from Schnakenberg & Fariss (2014) soaking up variance. The coefficient for the lagged human rights score is negative and statistically significant in the scarring model, indicating that AI is less likely to detect torture in countries with better human rights behavior in the immediate past; a good human rights record leads to less scrutiny by AI. The lagged human rights score has a negative coefficient in the clean model as well, although it is not statistically significant. Finally, country wealth has a positive impact on the probability of detecting both scarring and clean torture, although the coefficient is only significant for the scarring model. That covariates such as recent human rights practices and economic development have a smaller impact on AI's ability to produce allegations of clean torture is not surprising since clean torture is, by definition, conducted in a way that makes it more difficult to detect.

Fariss (2014, p. 297) documents that "over the past 35 years... [t]he standard of accountability used to assess state behaviors becomes more stringent as monitors look harder for abuse, look in more places for abuse, and classify more acts as abuse." As a result, projects that use the "naming and shaming' reports of human rights watchdogs as a source for content analysis of the rights violations of states and then examine the trend over time are engaged in a faulty practice: the values of the variables are not constant over time (see, aslo, Clark & Sikkink, 2013; Landman & Carvalho, 2009). More precisely, for any arbitrary ordinal (or other) scaling of "respect for rights" (or "abuse of rights") the level of activity that would be scored value k is not constant over time. Some might interpret this fact as causing subsantial problems for the ITT SA data, which are produced via content analysis

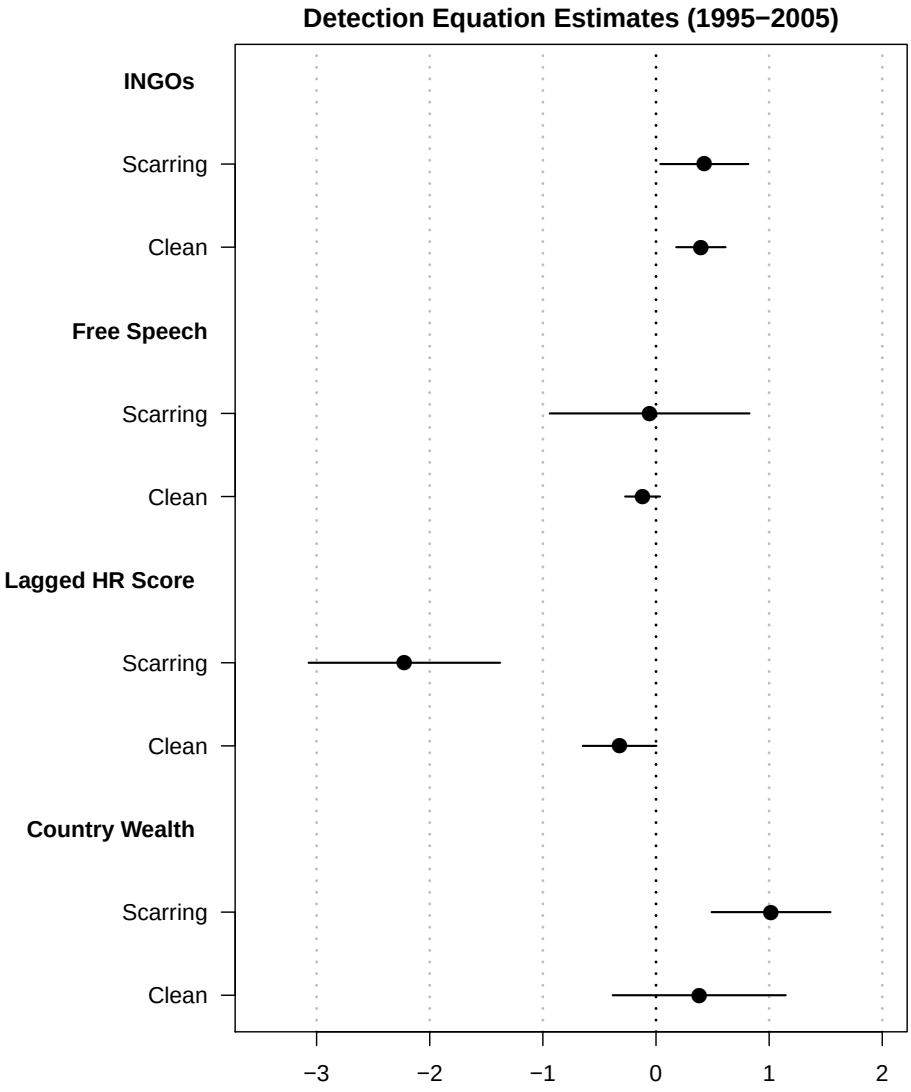

Figure 5: **Coefficient estimates from detection (logit) equation**. *The estimates are shown as dots, with 90% confidence intervals shown as lines. Where the line does not cross zero the coefficient is statistically significant at the* $\alpha = 0.10$ *level (two-tailed).* $N = 906$.

of AI's documents over the years of 1995-2005. They would be mistaken.[12]

Fariss (2014, pp. 309-10) reports that when he disaggregates all respect for all physical integrity rights into components the measure for torture is largely static,[13] noting that the

> lack of temporal change in the ITT data may reflect the relatively short period of coverage (1995–2005) or difference between Amnesty's Urgent Action Reports, upon which these data are [partly] based, and the annual report used bu the other data sources.

Thus, for our period of inquiry, Fariss finds that with respect to torture the standards are effectively static: the probability of coding the value k for ITT's CY Level of Torture variable does not change over time.[14]

That said, even if Fariss had found evidence of an increasing standard with respect to watchdogs' naming and shaming with respect to torture between 1995 and 2005, that would not impugn the analyses in our study. When the ITT project was conceived the PI's were aware of this problem and expressly designed the effort to address the fact that the true level of violations (1) can never be observed and (2) the undercount bias would vary across both countries and over time. In their funding proposal to the National Science Foundation, the PIs proposed to address the "unobservability / undercount bias" by coding what was observable—allegations—and using that information as input to a model of what

[12]Had we used the ITT Country Year (CY) variable "Level of Torture," then this issue would require consideration, as that variable is a standards measure that is subject to precisely the issue documented by Fariss (2014). As we are using the ITT SA data, however, the issue does not apply, as we explain below.

[13]See, especially, Figure 5.

[14]See, also, Figures 13 and 15 and Table 3 in the Appendix to Fariss. Further, please note that Fariss used the ITT CY's "Level of Torture" variable as input to his model, not the SA data used here. Nevertheless, in the interest of avoiding an endogeneity problem, we lag Fariss's measure one year before including it in our detection equation models.

was unobservable: state's (lack of) respect for right to freedom from torture.

In this study we use the undercount negative binomial regression model to accomplish this. Importantly, as described above, that statistical model permits us to model the probability of *observing and recording* a case of a torture event, as a function of covariates, across observations. That is, as the values of covariates change across observations, the probability of observing and recording a torture event changes as well (i.e., it varies across the values of the independent variables in the detection equation, which is to say, across countries and over time). Naturally, as with all regression models, the extent to which the undercount negative binomial regression model can represent the true undercount process depends upon how well it is specified (i.e., contains both the correct funtional form and the proper covariates). That is, the empirical model of the undercount process will be no better than our theory of that undercount process. We can also observe, however, that failure to model the undercount bias with a model that can account for a variable amount of bias across countries and over time will run afoul of the problem documented in Fariss (2014), though we submit that the shifting standard issue will be a relatively minor problem relative to the variation in undercount bias due to variable detection of violations that occur. Our submission is a conjecture, but one we would be very surprised to learn was inaccurate.

Further, please note that in broad strokes, the proposed approach is equivalent to the one pursued by Fariss (2014): using observable information to estimate a latent variable. The latent variable in our model is the probability that AI detects a torture event that has occurred. That probability is of central interest to anyone interested in using the al-

legations of watchdogs to study violations who is not willing to assume that the watchdog either observes all violations, or observes a constant undercount that does not vary across circumstances. As noted in the paper, we include a lagged value of the Fariss measure in our detection equation, thereby including as a regressor the state's previous level of respect for all physical integrity rights, corrected for the varying standards of rights monitors over time. The takeaway point is this: to suggest that the Fariss (2014) research impugns the use of the ITT data in this paper indicates a fundamental misunderstanding of both the Fariss (2014) article and the current effort.

4 Alternative Explanations & Robustness Checks

Both of our hypotheses are normatively unappealing. Democracy and elections are generally thought to be strongly and positively associated with normatively appealing outcomes, yet we show that elections are positively associated with scarring torture and powerful courts with clean torture. Churchill (1974) reminds us that democratic institutions are not necessarily linked to normatively appealing outcomes:

> Many forms of Government have been tried, and will be tried in this world of sin and woe. No one pretends that democracy is perfect or all–wise. Indeed, it has been said that democracy is the worst form of Government except all those other forms that have been tried from time to time...

Nevertheless, it is incumbent upon us as researchers to probe alternative accounts that might "explain away" the results reported above. These results will be available in the replication

dataset that we will post online upon publication.

To begin, elections and powerful courts are likely to be strongly positively correlated. When two independent variables are collinear in a regression, variances become inflated, and the sign of a variable's coefficient estimate can even become reversed (Wooldridge, 2003, 93-5, Kennedy, 2005, 86). In our data, the Pearson's r is 0.71. We provide a visual representation of their relationship in Figure 6. While these variables are strongly related, there are countries that hold competitive elections and have weak judiciaries, as well as those that do not hold elections and have strong judiciaries. If we define a weak/strong judiciary as one that scores below/above the mean of the Linzer-Staton measure, then our estimation sample includes 125 observations with competitive elections and weak judiciaries, and 41 observations with elections and strong judiciaries. This amounts to 18% of our sample ($N = 906$). Examples of countries with elections and weak judiciaries include Guatemala, Ukraine, and Indonesia since 1999. Countries with no competitive elections but strong judiciaries include Mexico the last three years the PRI were in power (1997-1999), Zambia, and Kuwait.

To establish that the signs on the coefficients for elections and the judicial power variables are not due to variance inflation induced by multicollinearity, we re-estimated the models reported in Figures three and four of the article, first excluding the judicial power variable, and then excluding the elections variable. The signs and significance for three of the four remain unchanged, but the positively signed estimate for judicial power in the clean equation becomes non-significant when we exclude elections from that equation. To summarize, none of the results reported in the study are due to degree of multi-collinearity

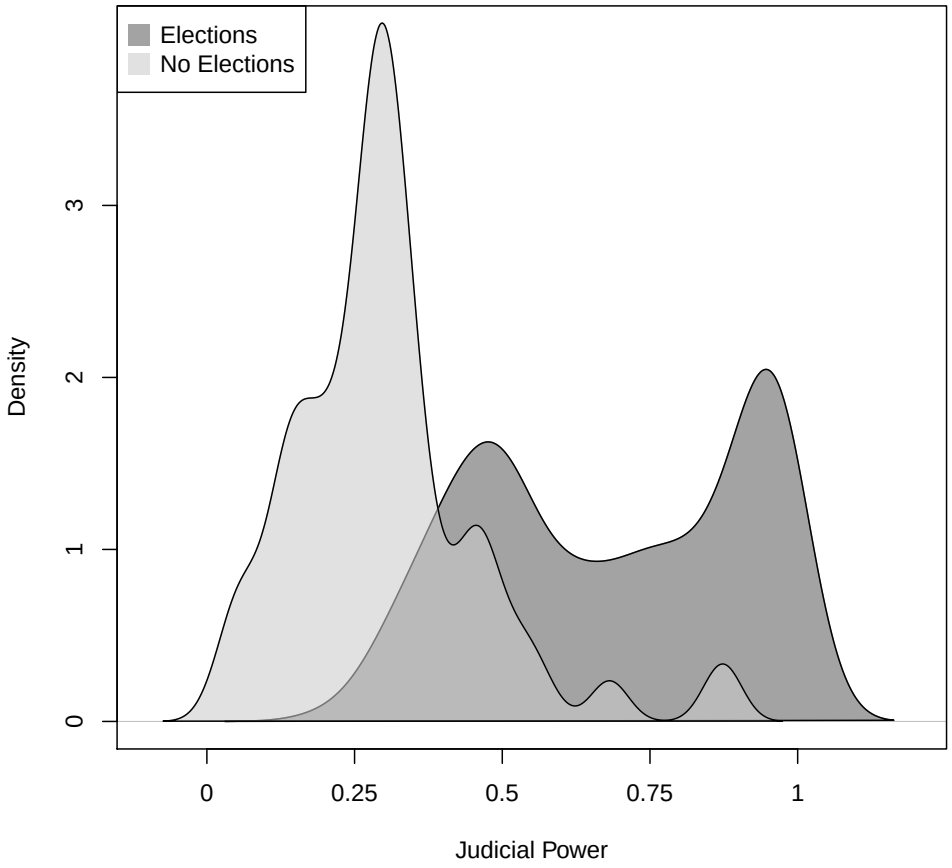

Figure 6: **Relationship between Electoral Contestation and Judicial Power, 1995-2005**. *The distribution of the judicial power score is shown separately for country-years that have a value of 1 on the competitive elections variable (shown in dark grey) and those that have a value of 0 on the competitive elections variable (shown in light grey).*

between elections and judicial power.

Second, we consider the possibility that cases with outlier values are influencing (i.e., biasing) the results. Perhaps AI has a tendency to shame particular democratic countries beyond that we capture with our detection equation, and those cases are influencing the results. To examine this possibility we identified, among the countries that hold elections, the country–years that produced the highest counts of both scarring and clean torture, and estimate additional models where these countries are (sequentially) omitted. The estimate for the elections variable was unchanged in all cases. The estimated coefficient for judicial power, however, becomes significant in the scarring equation when we excluded Romania, and becomes non-significant when we exclude either Israel or the US. The estimated value drops to 0.27 (se = 0.22) when Israel is left aside, and 0.36 (se = 0.22) if we hold out the US. In addition, a referee asked us to drop both Indonesia and Turkey from the sample. We did so, and again, the hypothesis tests were not affected.

Third, we establish that the results are unaffected by the exclusion or inclusion of several variables. To establish that the decision to use the 1,000 deaths threshold for our measure of civil war did not impact our findings, we re-estimated the models using the UCDP's 25 deaths threshold variable (Themnér & Wallensteen, 2013). We also estimated models for which we added two additional measures of the quality of information available to AI: the total number of "naming and shaming" documents issued by AI in the prior year that name the country (Meernik et al., 2012) and the the number of human rights organizations with an office in the country during that year (Smith & Wiest, 2005).[15] AI's "naming and

[15]The human rights organization data are available only through 2003, so the sample for those estimates

shaming" documents include Action Alerts, Briefings, and Press Releases. Human rights organizations are those listed in the *Yearbook of International Organizations* that had an office in the country and listed human rights as a topic on which they worked. For those models the results with respect to our hypotheses remain the same as those reported in the paper. Lastly, we dropped the freedom of expression measure from both the detection equation and the count equation, and the results did not change meaningfully.

Fourth, one might argue that a preferable specification of the statistical model would examine a ratio of the count of scarring and clean allegations. Indeed, in our very first empirical modeling for this project we adopted that approach, but learned that audiences responded poorly to that specification, consistently demanding to know what the results were for the alternative reported in the text above. A second reason persuaded us to abandon our initial approach and adopt the one reported in the body: we are unaware of an existing mixture model that would be appropriate to the task (i.e., we venture that mixing a beta regression model with a logit equation would be the place to start). As we are applied researchers and not metholodogists we chose to heed the advice we received from readers of our initial effort and pursue the models reported above. Nevertheless, while the estimates from a beta regression model "control for" rather than model the biased undercount problem, we have nonetheless estimated them and can report that the inferences we draw regarding H1 are not different using those models. The result for H2, however, is not sustained: judicial power still has a negative sign, but we are only able to reject the null with 84% confidence.[16]

is truncated by two years.

[16]The results from those models are included in a replication data set for this study.

To summarize, the results reported in the study are not due to multi-collinearity, nor are the results sensitive to the exclusion or inclusion of several variables. Finally, the support for hypothesis one is not fragile to the exclusion of cases with high values of scarring torture.

References

Achen, Christopher H (1986) *The statistical analysis of quasi-experiments*. Berkeley University of California Press.

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.

Churchill, Winston S (1974) Speech, house of commons, november 11, 1947. In: Robert Rhodes James (ed.) *Winston S. Churchill: His Complete Speeches, 1897–1963*. New York Facts on File , vol 7, 7566.

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.

Clark, Ann M & Kathryn Sikkink (2013) Information effects and human rights data: Is the good news about increased human rights information bad news for human rights measures? *Human Rights Quarterly* 35(3): 539–568.

Conrad, Courtenay R; Jillienne Haglund & Will H Moore (2013) Disaggregating torture allegations: Introducing the ill-treatment and torture (itt) country-year data. *International Studies Perspectives* 14(2): 199–220.

Conrad, Courtenay R; Jillienne Haglund & Will H Moore (2014) Torture allegations as events data: Introducing the ill-treatment and torture (itt) specific allegations data. *Journal of Peace Research* 51(3): 429–438.

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.

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

Feinstein, Jonathan S (1990) Detection controlled estimation. *Journal of Law and Economics* 33(1): 233–276.

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

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.

Kennedy, Peter E (2005) Oh no! i got the wrong sign! what should i do? *The Journal of Economic Education* 36(1): 77–92.

Landman, Todd & Edzia Carvalho (2009) *Measuring Human Rights* Taylor & Francis.

Linzer, Drew A & Jeffrey K Staton (2012) A measurement model for synthesizing multiple comparative indicators: The case of judicial independence. *Unpublished Manuscript*.

Long, J. S & Jeremy Freese (2001) *Regression Models for Categorical Dependent Variables Using Stata*. College Station Stata Press.

Meernik, James; Rosa Aloisi; Marsha Sowell & Angela Nichols (2012) The impact of human rights organizations on naming and shaming campaigns. *Journal of Conflict Resolution* 56(2): 233–256.

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

Ron, James (1997) Varying methods of state violence. *International Organization* 51(2): 275–300.

Schnakenberg, Keith E & Christopher J Fariss (2014) Dynamic patterns of human rights practices. *Political Science Research and Methods* 2(1): 1–31.

Smith, Jackie & Dawn Wiest (2005) The uneven geography of global civil society: National and global influences on transnational association. *Social Forces* 84(2): 621–652.

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

Staton, Jeffrey K (2010) *Judicial Power and Strategic Communication in Mexico*. New York Cambridge University Press.

Staton, Jeffrey K & Will H Moore (2011) Judicial power in domestic and international politics. *International Organization* 65(3): 553–587.

Themnér, Lotta & Peter Wallensteen (2013) Armed conflicts, 1946–2012. *Journal of Peace Research* 50(4): 509–521.

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

Wooldridge, Jeffrey M (2003) *Introductory Econometrics: A Modern Approach (2 ed.)*. Thompson/South-Western.