Rewarding Human Rights? Selective Aid Sanctions against Repressive States

Supporting Information

Rich Nielsen*

August 6, 2012

This web appendix provides additional information to support the analysis in "Rewarding Human Rights? Selective Aid Sanctions against Repressive States." Section 1 shows evidence that aid for social sectors and human rights is more likely to be delivered through NGOs than aid for economic sectors. Section 2 provides details about my operationalization of human rights coverage by the media. Section 3 presents the results of alternative specifications of the key statistical models.

1 How Much Aid Flows Through NGOs?

I argue that most donors impose aid sanctions by withdrawing aid for economic sectors, while simultaneously continuing to provide aid for social sectors and human rights. This is because economic aid is more fungible — and thus more valuable to recipient governments — than social aid or human rights aid. Direct control of foreign aid funds is the most obvious way in which aid flows are fungible; I argue that recipient governments are more likely to directly control aid to economic sectors. In contrast, aid for social sectors and human rights

*Ph.D. Candidate, Department of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138. Email: nielsen.rich@gmail.com. I appreciate the comments of Muhammet Bas, Sarah Bermeo, Oliver Bevan, Sheena Chestnut, Andrew Coe, Zachary Davis, Michael Findley, Jay Goodliffe, Darren Hawkins, Iain Johnston, Josh Loud, Vipin Narang, Eric Neumayer, Rebecca Nielsen, Daniel Nielson, Chris O'Keefe, Jonathan Renshon, Beth Simmons, Jane Vaynman, Erik Voeten, and three anonymous reviewers. This research was supported by a National Science Foundation Graduate Research Fellowship. Replication materials and a Supporting Information appendix are available online at `http://dvn.iq.harvard.edu/dvn/dv/rnielsen` and `http://www.isanet.org/pubs/data-archive.html/`.

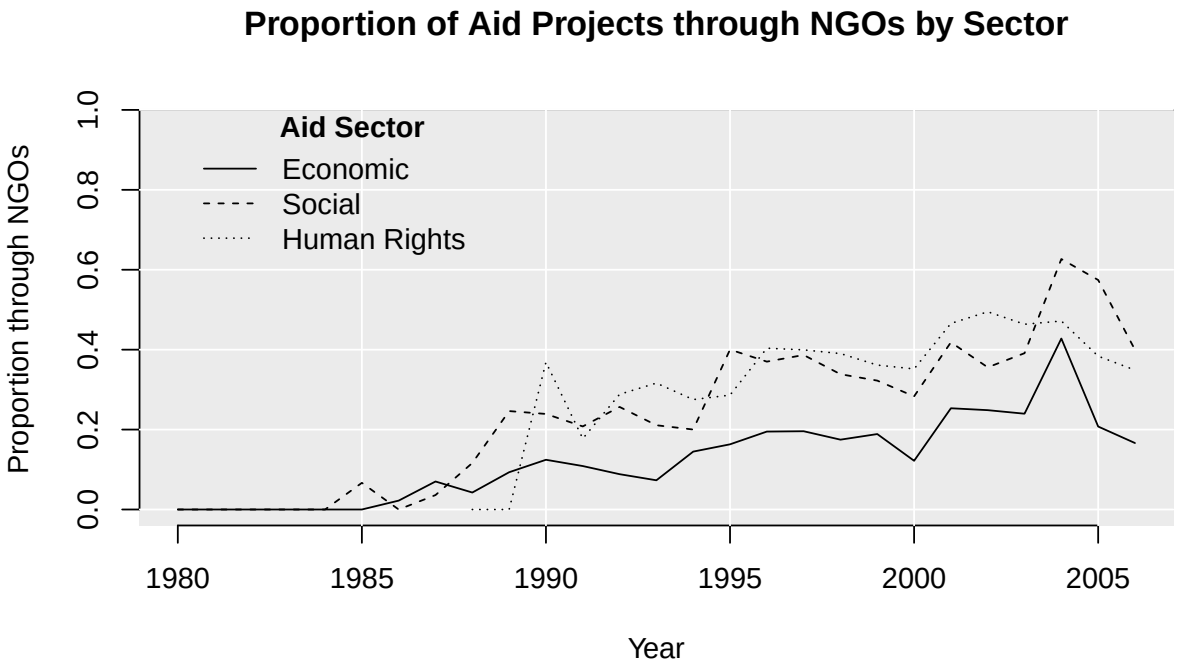

Figure 1: *This figure shows data on the delivery of aid projects through NGOs by sector over time. The coding of "non-governmental" delivery is intentionally conservative: if I was unable to determine how an aid project was delivered, I assumed that it was delivered through the recipient government.*

are more likely than economic aid to be delivered through third party NGOs that can keep the recipient government from seizing the funds.

This is a crucial point — if these types of aid are not delivered through NGOs at higher rates than other projects, my conclusions about the differences between aid sectors could be problematic. To show that these projects are implemented by NGOs at higher rates, I first collected project-level data on the "channel of delivery" as reported to the OECD by donors. Unfortunately, this data is not reported consistently by the various donors; in most cases, they simply list the name of an organization. To determine whether project funds were delivered through government or non-government channels, I coded 114,018 entries as "governmental" or "non-governmental." I obtained information on many organizations by searching for them on the Internet and reading their websites to make a determination. For the 43,000 entries where I could not make a determination (including many entries of missing data), I made the conservative assumption that the funds were delivered through the recipient government.

Supporting my theoretical arguments, I find that aid for human rights and social sectors are more likely to flow through NGOs than aid for economic sectors. From 1990 to 2005,

between 35 and 50 percent of human rights aid was delivered through NGOs (note that there was very little human rights aid prior to 1990). Over this same time period, about 35 percent of social aid projects were delivered through NGOs compared to only 20 percent of economic aid projects. These numbers are likely to understate the true levels of NGO use by donors because of the entities that I was unable to code.

2 Measuring Media Coverage of Rights Abuses

In the paper, I find that aid donors are more likely to impose sanctions for human rights violations that receive media attention. This section describes how I measure media coverage of human rights violations using news articles from the *New York Times*.

I focus on news coverage by the *New York Times* because it is accessible, prominent in the United States, and thorough. Ideally, I would measure media coverage of rights abuses within each donor country because some abuses may be reported by news media in some donors but not others. However, collecting and classifying news reports by the leading newspapers in each of 17 donor countries was too time consuming so I assume that media coverage of rights abuses is fairly constant across OECD donors. If rights violations are reported by a prominent newspaper in the US, they are probably also reported in Europe and Japan.

I used two methods for measuring news coverage of human rights by the *New York Times*. The first (featured in the text) uses a simple set of LexisNexis searches. The second expands on these searchers using automated text analysis methods to try to eliminate false positives and negatives. Ultimately, I find that both are measuring the same thing (the results correlated at 0.98).

2.1 Basic LexisNexis Searches

To measure coverage of human rights in the *New York Times*, I used *LexisNexis* and searched (using the example of Uganda) "human right! w/25 Uganda", which returns the citation to every article that has a variant of the phrase "human right(s)" occurring within 25 words of the phrase "Uganda". This rudimentary type of text analysis was necessary because LexisNexis protects its articles vigorously, making it difficult to download the full text of these articles and use more sophisticated content analysis techniques. Obviously, this technique produces a number of false positives (articles that are not really about human rights in Uganda but that fit these search criteria) and false negatives (articles that are about human rights in Uganda but that do not fit these search criteria). This is inevitable, but to try to minimize it, I experimented with different separation distances — 5, 10, 25, 50, and 100 words — between the phrase "human right" and the country name. To assess which separation distance produced the best balance of false positives and negatives, I hand-coded a set of New York Times articles about Uganda (for the years 2003-2005) that mentioned "human right(s)" anywhere in the article. I then compared the number of false positives and negatives that each separation distance produced. As expected, I found that the longest distance (100 words) had the fewest false negatives, but I also found that a distance of 25

words (used in the paper), kept most of the correctly classified articles while discarding a number of the false positives. Shorter distances such as 5 and 10 words began to produce unacceptably large numbers of false negatives.

After collecting the citations for the articles returned by my searches,[1] I parsed the citations using the regular expression functions in `R` and created a dataset of the number of New York Times articles mentioning "human rights" in close context with each recipient country.

The counts of New York Times coverage of abuses are correlated with actual rights violations as coded by the CIRI physical integrity index, but the two are hardly identical (the Pearson correlation is 0.46). Figure 1 shows the standardized level of physical integrity violations over time and the standardized number of New York Times articles about abuses for four cases.[2]

2.2 Automated Content Analysis of Articles from LexisNexis

I then undertook further efforts to avoid the possibility that false positives and negatives were biasing the results with respect to media coverage of human rights. My simple *LexisNexis* searches have clear limitations, the most important being that I did not collect the full text of the articles I counted so there was no way to see whether false positives were a serious problem without recollecting the data. I went back to *LexisNexis* and re-downloaded the results of the same searches, this time keeping the full text of each article. I then took a random sample of 400 news articles to see how many were false positives. Very few mentions of human rights related to praise for human rights improvement (about 1 in 100). However, when I asked a more restrictive coding question: "Does this article specifically mention human rights violations, past or present, in the country?", I found that only 240 out of 400 sampled articles met this requirement. This means that there is likely to be substantial measurement error in the original NYT variable.

I first tried to do a complete hand-coding of the 30,789 news articles but stopped because this would have taken roughly 500 hours (60 workdays). I then used the 400 articles I coded by hand to build a statistical model to automate the coding of the remaining articles.[3] Unfortunately, the model was only able to achieve 60% accuracy at correctly matching my hand-coding on the 400 articles in the training set (this is actually quite good for an automated coding task of this complexity). Specifically, the coding task is hard because the statistical model is trying to estimate both the context in which human rights are mentioned (criticism or praise) *and* whether the country being criticized or praised is the country whose name was in the original search query. My inability to get better than 60% accuracy with statistical text methods suggests that there is no systematic component to the errors made by my initial text searches. The machine coded variable has a correlation of 0.98 with the

[1] These searches were the time-consuming part. Because of the terms of use of LexisNexis, I had to do a search for each of 118 recipient countries by hand.

[2] They are standardized so that they both range from zero to one.

[3] I tried both support vector machines and random forests. They each gave essentially identical performance.

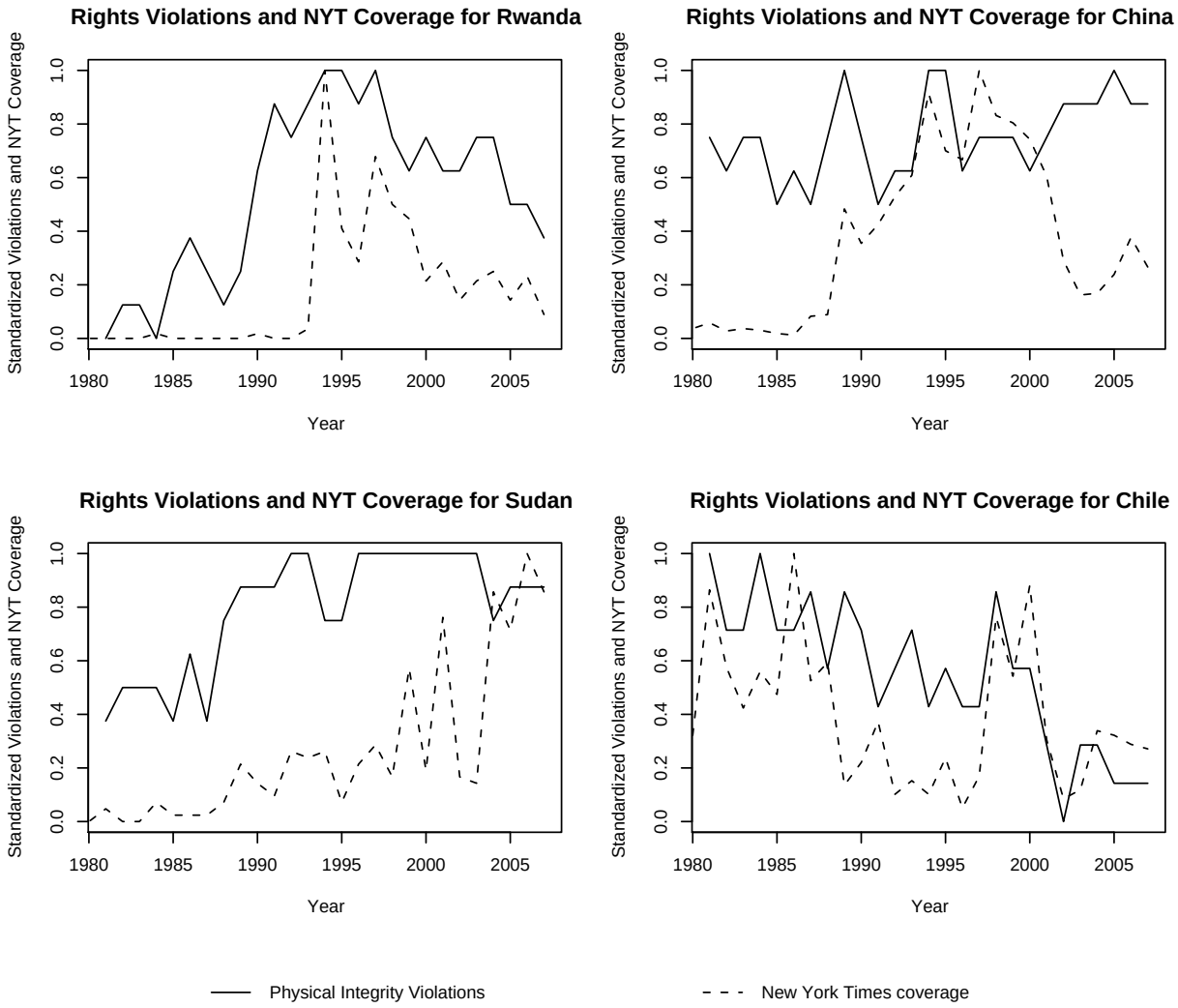

Figure 2: *Physical integrity violations (standardized to range from zero to one) and New York Times Coverage of Human Rights (standardized to range from zero to one) for four cases: Rwanda, China, Sudan, and Chile.*

simple counts from the LexisNexis searches. Thus, the original variable from the simple LexisNexis searches seems to be measured with noise, but without systematic bias.

Table 2.2 presents a model with this machine coded NYT variable.

Dependent Variable: Economic Aid	Estimate	Standard Error
ln(Machine Coded NYT Coverage)	-0.019	0.054
ln(Machine Coded NYT Coverage) × Violations	-0.021*	0.0098
Donor Ratification	-0.30	0.20
Donor Ratification × Violations	0.079	0.040
Donor Rights Protection	-0.11*	0.058
Donor Rights × Violations	0.013	0.013
Ally Neighbor	0.71**	0.22
Ally Neighbor × Violations	-0.082*	0.036
Alliance	0.06	0.22
Alliance × Violations	0.095*	0.045
UN Friend	-0.92**	0.23
UN Friend × Violations	0.15**	0.046
ln(Refugees)	0.045	0.020
ln(Refugees) × Violations	-0.0046	0.0036
Violations	-0.022*	0.098
Democracy	0.030**	0.005
Lagged Aid	0.45**	0.009
Global Aid Flows	0.80**	0.027
ln(GDP per capita)	-0.57**	0.066
ln(Population)	0.32*	0.04
ln(Trade)	-0.073**	0.009
Former Colony	1.4**	0.32
Socialist	-0.75**	0.16
post-Cold War	-0.027	0.083
post-Cold War × Socialist	1.00**	0.11
Post-Cold War × Violations	0.024	0.014
War	-0.066	0.067
post-2001	0.21**	0.056
Observations	42,367	
Left-censored observations	24,733	
Number of dyads	2,385	
Log-likelihood	-52984.85	

Table 1: *Area dummies and a constant term were included in the model but omitted from the table. ** p<.01, * p<0.05.*

3 Alternative Statistical Models

Like most studies using statistical analysis, I examined many more models than I am able to conveniently present in the text. This section reports the results of 84 of these models.

Each model or set of models is numbered consistently in the subsections and figures that follow. To quickly summarize the results of all 84 models, I plot the point estimate and 95% confidence interval of the important coefficient(s) for each model in Figures 2-4. The coefficients of control variables are not reported to save space. In each numbered section, I describe the models and general results. Full replication materials are available so that researchers interested in specific models can easily reproduce them.[4]

Figure 2 shows a series of alternative specifications of the models of sector aid allocation. My theory suggests that aggregate aid may have a mixed relationship with human rights violations, aid for economic sectors should have a negative relationship with violations, social sector aid should have no relationship, and human rights aid should have a positive relationship. At the top of Figure 2, I show the point estimates for the coefficient on the variable *physical integrity violations* obtained from the original models show in table 1 of the paper (horizontal bars are 95% confidence intervals). Then, for comparison, I plot the coefficient on the same variable, physical integrity violations, in a number of reasonable alternative specifications. Each specification is labeled with a number that corresponds with an explanation in the numbered paragraphs of the next few pages. Some of these coefficients are not in the same units as the original models so direct comparison is not always useful. However, it is always useful to note whether each coefficient is to the left or right of zero and whether its confidence bands cover zero. I expected the coefficient on rights violations to be negative for economic aid, zero for social aid, and positive for human rights aid.

Figures 3-4 show a series of alternative specifications testing why donors sanction rights violations. In each model, the relevant coefficient for comparison to the original models (again plotted above) is to look at the direct, magnitude, and statistical significance of the interaction term(s) included in each model.

Most substantive finding discussed in the text holds up in most of the alternative specifications and in the cases where there are contrary findings, there are clear statistical reasons why the results might be different. These are discussed along with each model in the following pages.

[4]Replication materials are available at `http://dvn.iq.harvard.edu/dvn/dv/rnielsen` and `http://www.isanet.org/pubs/data-archive.html/`

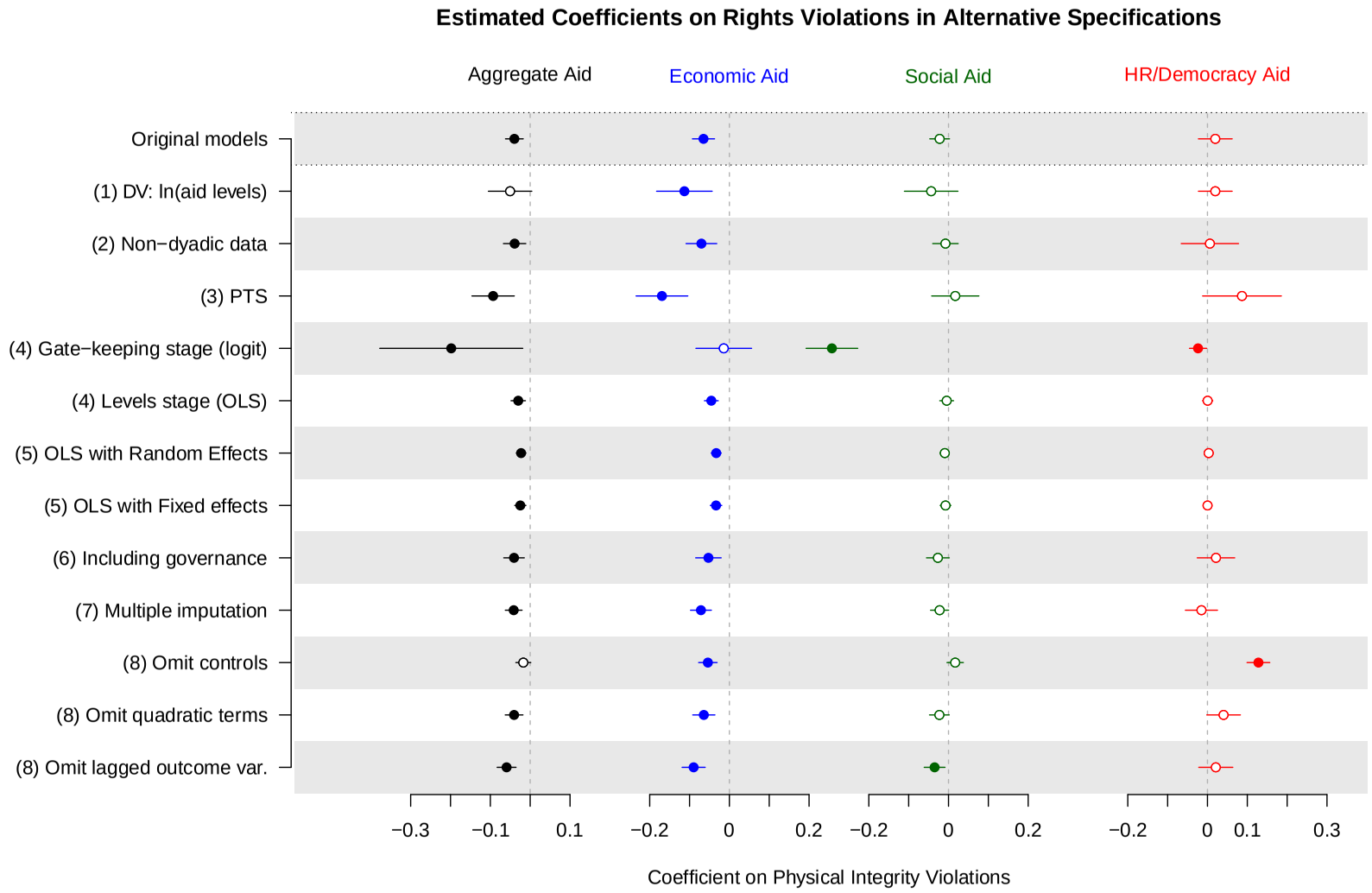

Figure 3: *Coefficients and 95% confidence intervals for alternative specifications. Filled points are statistically significant at the 95 percent level. Numbers in parentheses refer to the numbered list below where there is information about each model.*

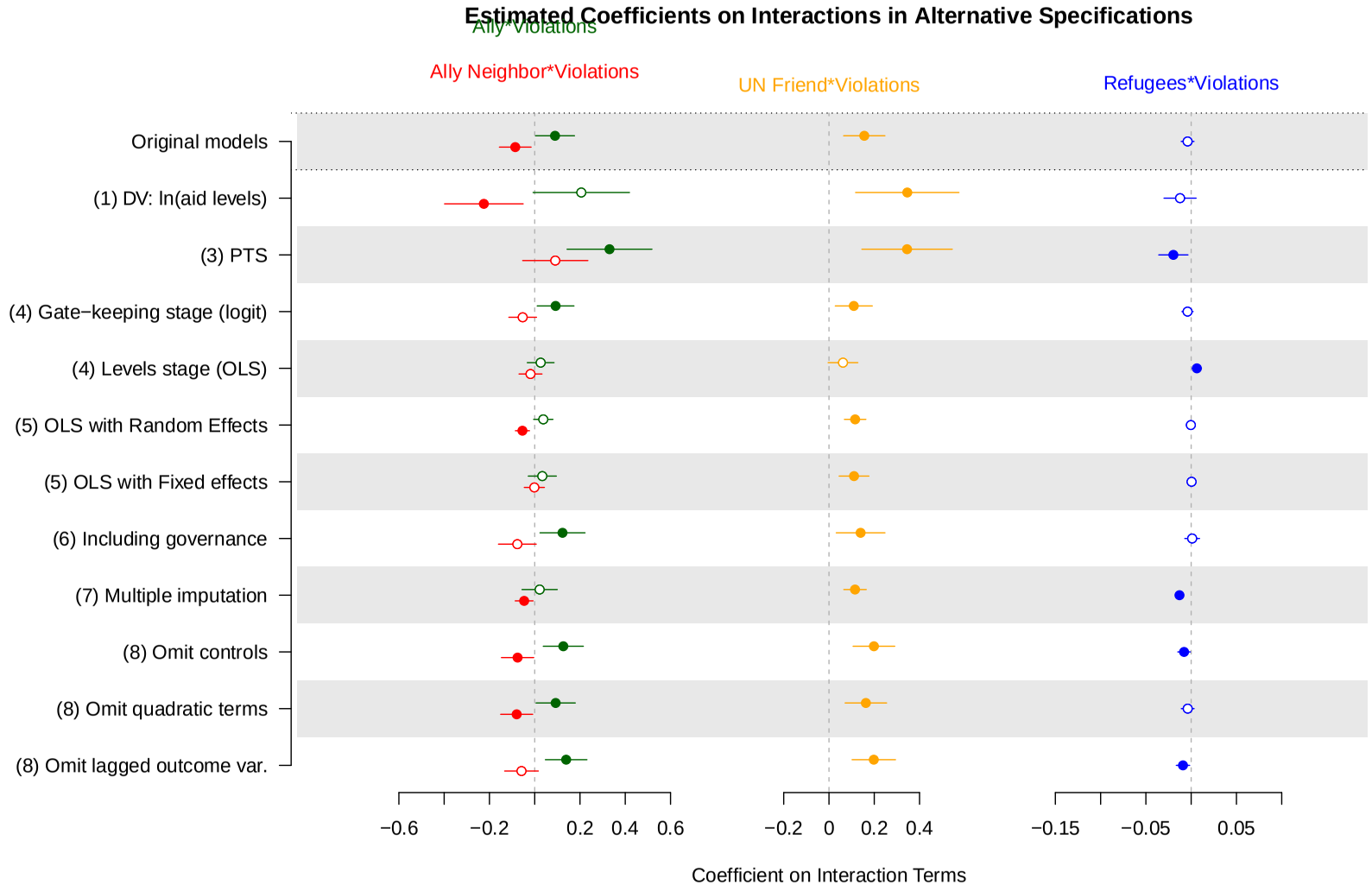

Figure 4: *Coefficients and 95% confidence intervals for alternative specifications. Filled points are statistically significant at the 95 percent level. Numbers in parentheses refer to the numbered list below where there is information about each model.*

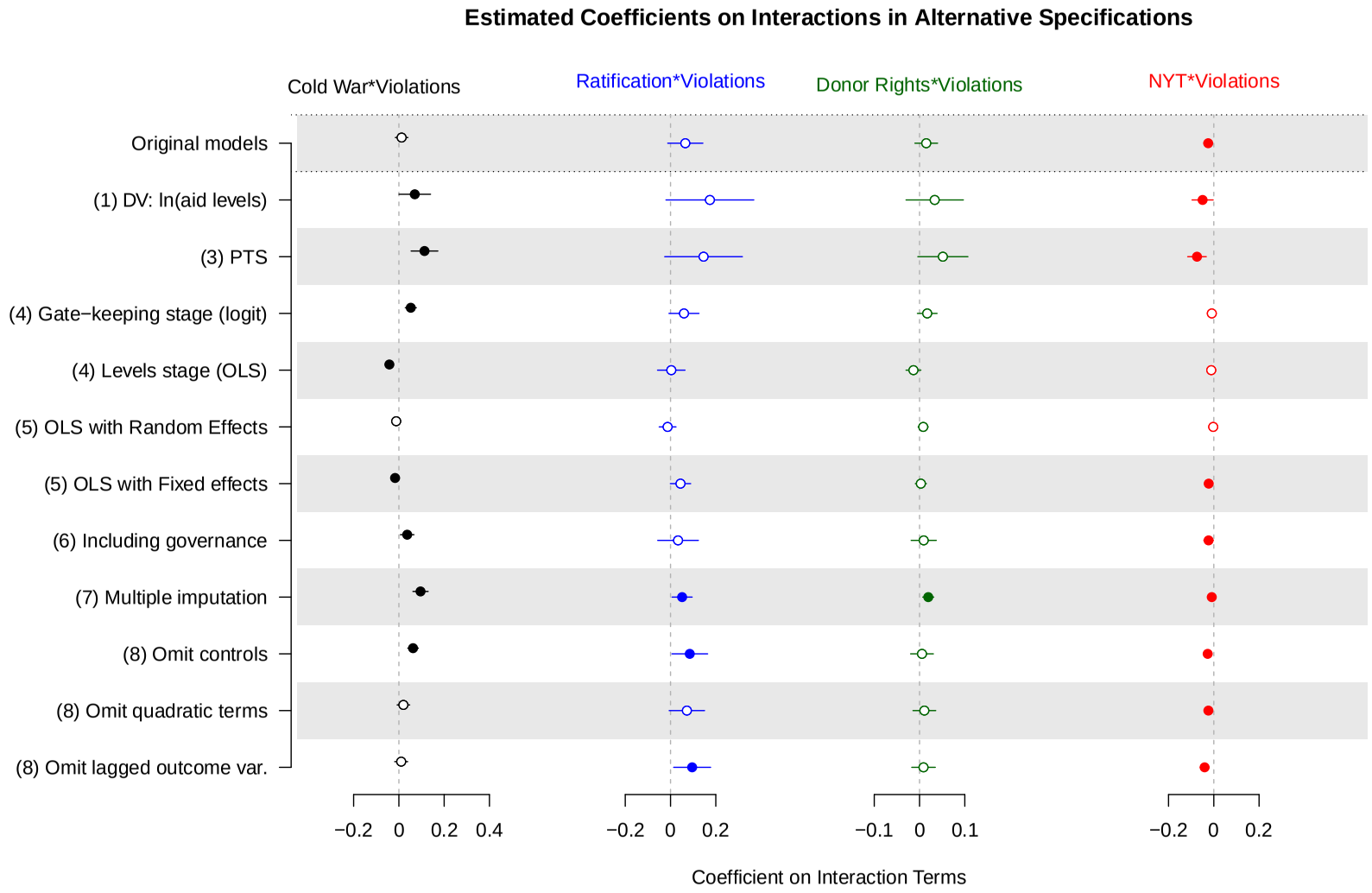

Figure 5: *Coefficients and 95% confidence intervals for alternative specifications. Filled points are statistically significant at the 95 percent level. Numbers in parentheses refer to the numbered list below where there is information about each model.*

1. **Robust, clustered standard errors:** The data display heteroskedasticy and serial correlation according to standard tests, meaning that estimating a pooled Tobit model without additional structure is likely to lead to incorrect standard errors (and potential bias in most quantities of interest, see King and Roberts (N.d.)). There are two broad approaches to fixing these problems: (1) model the underlying features of the data that create heteroskedasticity and serial correlation, or (2) estimate the original model but try to fix the standard errors after the fact with Huber-White-style errors (Primo, Jacobsmeier and Milyo, 2007). In the paper, I use hierarchical modeling (dyad random effects) because I prefer to directly model the underlying dyad heterogeneity generating the statistical issues rather than patch up the standard errors. However, the results are generally the same when I re-estimate the model with robust standard errors, clustered by dyad. The only changes of note are that the coefficients on the interactions *Ally*$\times$*Violations* and *ln(NYT Coverage)*$\times$*Violations* lose statistical significance, but retain the correct sign. Since heteroskedasticity and serial correlation can bias model estimates, this difference between the models with random effects and the models with robust errors (but no random effects) is not particularly troubling. The model that fixes the problems statistically rather than patching up the standard errors is almost always preferable (King and Roberts, N.d.).

2. **Alternative transformations of foreign aid:** The dependent variables in the primary specifications are measured as the natural log of aid per capita. I re-estimate the models with the dependent variables measured as the natural log of aid levels. Note that coefficients are incomparable to the coefficients of the original models. In general, the results are very similar to those presented in the paper. The only important changes to the interpretation would be that aid sanctions seem to be significantly less severe after 1991 and the decreased sanctions enjoyed by donors' allies are now only marginally statistically significant (but still quite large in magnitude).

   Another possible transformation is to difference receipt of foreign aid from one year to the next. To explore this possibility, I re-estimate the models presented in the paper as error correction models of the form $\Delta Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 \Delta X_t + \beta_1 X_{t-1} + \varepsilon_t$ (De Boef and Keele, 2008, see table 2). These models allow us to estimate both the short-term (same-year) effect of human rights violations on aid as well as a the long-term effect. The dynamics of aid suggest that any effects of human rights will be one or two years downstream, meaning that the "long-term" effect is our primary quantity of interest in these models. The results of these models are very similar patterns to those reported the paper.

3. **Non-dyadic data:** There is some reason to be skeptical of results obtained from very large dyadic datasets because large numbers of dependent observations (dyad-years are dependent within the dyad and probably the year, possibly even after conditioning on the dyad random effects). I use dyadic data because my theory requires dyadic testing. However, it is possible to estimate the models of sector aid allocation with non-dyadic, time-series cross-sectional data, where the unit of observation is the recipient

year and the dependent variable is aid for economic, social, and human rights sectors summed across donors. The results of these models are virtually identical to the findings presented in the paper.

4. **Alternative measures of human rights violations:** In the primary specifications I use the CIRI measure of physical integrity rights (Cingranelli and Richards, 2006). These models use an alternative measure known as the Political Terror Scale (PTS) (McCann and Gibney, 1996) that is also coded from State Department and Amnesty International reports. For a paper comparing these two scales, see Wood and Gibney (2010). Using the PTS changes two key results slightly. First, the interaction of refugee flows with *Violations* becomes significant and negative. Second, donors' allies are still exempt from aid sanctions but the neighbors of donor allies do not face harsher sanctions than other states in this model (the coefficient on the interaction is insignificant).

5. **Separate estimation of "gate-keeping" and "levels" stages of aid allocation:** Because of its ease of implementation and clean theoretical interpretation, many studies estimate the effects of rights violations on aid allocation by first estimating a logistic regression model predicting which potential recipients actually receive aid and then estimating a linear regression predicting how much aid recipients get (conditional on receiving aid). This approach has statistical drawbacks — the coefficients in the second stage are likely to be biased without difficult corrections.

   In general, other studies that have used separate estimation of "gate-keeping" and "levels" models find at least a few odd results that are difficult to make sense of theoretically. I find the same: for example, at the gate-keeping stage, donors seemed to sanction more *during* the Cold War but at the levels stage, donors seemed to sanction more *after* the Cold War. I also find odd inconsistencies in the responsiveness of different sectors of aid to violations: at the levels stage, I find similar results to those reported in the paper but at the gate-keeping stage, I find that economic aid is *not* sensitive to violations while social aid *is*, a puzzling result that is not replicated in any other model. Strategic factors such as alliances and UN voting similarity have their expected effects in the gate-keeping stage but lose statistical significance (while keeping the correct signs) at the levels stage.

   Ultimately, I do not place much faith in the two-stage modeling process because the statistical models are known to be biased and seem more aimed at modeling a stylized and overly simplistic story about how aid is allocated than modeling the actual aid allocation data. Statistical theory suggests that the results I report in the paper are more reliable.

6. **OLS with random and fixed effects:** In these models, I use OLS with random effects and, alternatively, fixed effects on the entire sample (as opposed to the conditional sample of states that received some aid). These models reduce the magnitude of

some coefficients. However, there are good reasons to expect that OLS estimates are biased if there is a large amount of censoring (zeros) in the outcome data.

Green et al (2001) argue that fixed effects should be included in dyadic time-series cross-sectional models of international data. Their argument is controversial (Beck and Katz, 2001) and fixed effects are known to be somewhat biased for Tobit models (Honore, 1992). I am unable to estimate dyadic fixed effects with the original Tobit models because the optimization routine fails to converge — estimating 2,430 country-level intercepts is taxing on the data. I estimate fixed effects OLS models, but these are subject to the criticisms of OLS for censored data discussed in the paper.

The results from OLS with random effects are very similar to the results from Tobit with random effects in the original models. The OLS with fixed effects gives a different result for alliances, suggesting that donors' alliances are not very important for decreasing sanctions (although similarity of UN voting is).

7. **Including measures of "good governance":** Bermeo (2007) shows that general measures of "good governance" (such as corruption indicators) are strong predictors of economic aid but have less of an impact on social aid. I check to make sure that my findings are not affected by including her measure of governance, the corruption index coded by the ICRG (2005). The relevant coefficients remain similar across most models indicating that my findings are not the spurious result of donors' attempts to reward "good governance" generally.

8. **Multiple imputation of missing data:** Data are missing for some 18,408 dyad-years. I define the full dataset as every independent state between 1981 and 2004 and multiply impute the missing data five times using Amelia software (Honaker, King and Blackwell, 2005–2008). I then re-estimate the models of each imputed dataset, averaging the coefficients and adjusting the standard errors accordingly. The fully rectangular dataset now includes 63,987 observations and 3,045 dyads, up from 44,277 and 2,366, respectively, but these are generally micro-states that do not receive much aid. Only one key result is called into question when I use multiple imputation — it is no longer clear that donors' military allies are exempt from sanctions. Interestingly, the coefficient on *ln(Refugees)* $\times$ *Violations* is now negative and significant suggesting that refugee spillovers may influence aid sanctions once we account for missing data.

9. **Reduced form models:** Achen (2005) argues that regressions with large numbers of explanatory variables — "garbage can regressions" — can obscure important relationships in the data and lead to spurious findings. I re-estimate the main models without most of the control variables, leaving only the lag of sector aid per capita, the lag of global sector aid, and the dyad-level random effects. The findings are surprisingly robust to this drastic change in the model and the general conclusions remain unchanged.

10. **Including/Excluding particular covariates:**

I explored the possibility that some covariates might have nonlinear relationships with Aid using a series of Generalized Additive Models (GAMs): a form of generalized linear model in which some variables are specified to be fitted using a series of smoothing splines (Hastie and Tibshirani, 1990). There is some evidence in the data that *ln(GDP per capita)*, *ln(Population)*, and *ln(Trade)* have somewhat non-linear relationships with aid allocation. Ultimately, I did not include quadratic — $ln(GDP\ per\ capita)^2$, $ln(Population)^2$, and $ln(Trade)^2$ — in the model for a few reasons, including possible collinearity, but adding these terms does not change the results.

There has been some debate about the appropriateness of lagged dependent variables (Keele and Kelly, 2006). Although the lagged dependent variables have a strong theoretical justification because of the bureaucratic "stickiness" of aid, my findings are robust to the exclusion of the lagged dependent variables.

Egypt and Israel were both recipients of exceptionally large amounts of foreign aid from the United States during the period of study, although much of it was military aid and thus falls outside the scope of my analysis. Still, some researchers worry that these two outlying cases might bias regressions results, especially because the reasons for which they received so much aid are idiosyncratic (essentially, to maintain peace). All of the results presented in the paper hold when I include indicators for the US-Israel and US-Egypt dyads, so these exceptional cases are not driving the result.

I include Global Aid Flows as a way to account for changes in aid flows to particular recipients that are just a result of changes in the overall generosity of donors (perhaps aid to all countries decreases because of hard times in donor countries). Omitting this control variable does not substantively change the results.

References

Achen, Christopher H. 2005. "Lets Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22:327–339.

Beck, Nathaniel and Jonathan N. Katz. 2001. "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon." *International Organization* 55 (2):487–495.

Bermeo, Sarah Blodgett. 2007. "Foreign Aid, Foreign Policy, and Development Sector Allocation in Bilateral Aid.".

Cingranelli, David L. and David L. Richards. 2006. "The Cingranelli-Richards (CIRI) Human Rights Dataset.". `http://ciri.binghamton.edu/` [Accessed November 1, 2006].

De Boef, Suzanna and Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52 (1):184–200.

Green, Donald P., Soo Yeon Kim and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55 (2):441–468.

Hastie, Trevor and Robert Tibshirani. 1990. *Generalized Additive Models*. New York: Chapman and Hall.

Honaker, James, Gary King and Matthew Blackwell. 2005–2008. "Amelia II: A Program for Missing Data.". Available at http://gking.harvard.edu/amelia (September 11, 2008).

Honore, Bo E. 1992. "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects." *Econometrica* 60 (3):533–565.

ICRG. 2005. "International Country Risk Guide." Electronic database.

Keele, Luke and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14 (2):186–205.

King, Gary and Margaret Roberts. N.d. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix." Unpublished manuscript. `http://gking.harvard.edu/gking/files/robust.pdf`.

McCann, James A. and Mark Gibney. 1996. *An Overview of Political Terror in the Developing World, 1980-1991*.

Primo, David M., Matthew L. Jacobsmeier and Jeffrey Milyo. 2007. "Estimating the Impact of State Policies and Institutions with Mixed-Level Data." *State Politics and Policy Quarterly* 7 (4):446–459.

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