Supplemental Appendix for The Prevalence and Severity of Underreporting Bias in Machine and Human Coded Data*

Benjamin E. Bagozzi, Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim, Agustin Palao Mendizabal & Carly Potz-Nielsen

Contents

*This research is supported by NSF Grant Number SBE-SMA-1539302.

Overview

In this supplementary material, we first provide a brief summary of the likelihood function for the Cook et al. two-source under-misclassification model, which we discuss in detail in our main Research Note. We then present our aforementioned Monte Carlo replications and extensions of Cook et al. (2017). Following this Monte Carlo section, we provide extended discussions and analyses of (i) our African repression application and (ii) our Colombian human rights violation (HRV) application. In each of these application sections, we first describe our data and aggregation decisions in detail, before presenting a series of bivariate comparisons and summary statistics. Following these bivaraite comparisons, each application then applies and interprets Cook et al.'s multi-source models, before finally validating the results obtained from these models with an external, gold standard records (GSRs). Lastly, we present the formulas for the classification statistics that we employ within our extended application discussions.

Cook et al. Likelihood

Recall that the joint probability statement for Cook et al.'s two-source under-misclassification model (as reported in our main Research Note) is:

$$\begin{aligned}
\Pr(\mathbf{Y_1}=0,\mathbf{Y_2}=0|\mathbf{X},\mathbf{Z_1},\mathbf{Z_2}) &= [1-F(\mathbf{X},\beta)] + \alpha_1(\mathbf{X},\mathbf{Z_1})\alpha_2(\mathbf{X},\mathbf{Z_2})F(\mathbf{X},\beta);\\
\Pr(\mathbf{Y_1}=0,\mathbf{Y_2}=1|\mathbf{X},\mathbf{Z_1},\mathbf{Z_2}) &= \alpha_1(\mathbf{X},\mathbf{Z_1})\left[1-\alpha_2(\mathbf{X},\mathbf{Z_2})\right]F(\mathbf{X},\beta);\\
\Pr(\mathbf{Y_1}=1,\mathbf{Y_2}=0|\mathbf{X},\mathbf{Z_1},\mathbf{Z_2}) &= \left[1-\alpha_1(\mathbf{X},\mathbf{Z_1})\right]\alpha_2(\mathbf{X},\mathbf{Z_2})F(\mathbf{X},\beta);\\
\Pr(\mathbf{Y_1}=1,\mathbf{Y_2}=1|\mathbf{X},\mathbf{Z_1},\mathbf{Z_2}) &= \left[1-\alpha_1(\mathbf{X},\mathbf{Z_1})\right]\left[1-\alpha_2(\mathbf{X},\mathbf{Z_2})\right]F(\mathbf{X},\beta)
\end{aligned} \tag{A.1}$$

This statement has the corresponding likelihood function (across N observations) of:

$$\begin{aligned}
\mathcal{L}(\beta,\eta_1,\eta_2) = \prod &\left\{[1-F(\mathbf{X},\beta)] + \alpha_1(\mathbf{X},\mathbf{Z_1},\eta_1)\alpha_2(\mathbf{X},\mathbf{Z_2},\eta_2)F(\mathbf{X},\beta)\right\}^{(1-\mathbf{Y_1})(1-\mathbf{Y_2})}\\
&\times\left\{\alpha_1(\mathbf{X},\mathbf{Z_1},\eta_1)\left[1-\alpha_2(\mathbf{X},\mathbf{Z_2},\eta_2)\right]F(\mathbf{X},\beta)\right\}^{(1-\mathbf{Y_1})\mathbf{Y_2}}\\
&\times\left\{\left[1-\alpha_1(\mathbf{X},\mathbf{Z_1},\eta_1)\right]\alpha_2(\mathbf{X},\mathbf{Z_2},\eta_2)F(\mathbf{X},\beta)\right\}^{\mathbf{Y_1}(1-\mathbf{Y_2})}\\
&\times\left\{\left[1-\alpha_1(\mathbf{X},\mathbf{Z_1},\eta_1)\right]\left[1-\alpha_2(\mathbf{X},\mathbf{Z_2},\eta_2)\right]F(\mathbf{X},\beta)\right\}^{\mathbf{Y_1}\mathbf{Y_2}}
\end{aligned} \tag{A.2}$$

Note that the likelihood in Equation A.2 is a slightly modified version of the likelihood reported in Cook et al. (2017, 228). Specifically, we have corrected the ordering of the exponents within the first and fourth lines of the likelihood's original presentation in Cook et al. (2017, 228) to accurately match the joint probability statements in Equation A.1.

Monte Carlo Extensions

This section replicates and extends the Monte Carlo experiments found in Cook et al. (2017). The original Monte Carlo experiments performed by Cook et al. (2017) primarily compare their multi-source misclassification models[1] to the following plausible alternatives:

[1] Specifically, the two-source constant misclassification-probability version of the Cook et al. estimator and the two-source version of Cook et al.'s estimator that includes covariates in the corresponding misclassification

a naive probit model, a Hausman (i.e., single/collapsed source) misclassification estimator with constant probabilities in the misclassification stage, and a Hausman misclassification estimator with covariates included in the misclassification stage (Hausman, Abrevaya and Scott-Morton, 1998). In each of these cases, Cook et al. examined a binary dependent variable with two reporting sources, wherein each reporting source exhibited moderate levels of underreporting. We extend these experiments below for two plausible situations that are likely to arise in applications of the Cook et al. model to political event data: (i) severe underreporting in *both* reporting sources analyzed and (ii) severe underreporting bias in one reporting source but very low underreporting bias in a second reporting source.

Extension 1

Extension 1 examines the case of severe underreporting bias within both (i.e., two) reporting sources. Here, we first consider Experiment 1 from the simulations performed by Cook et al. (2017). In Experiment 1, the authors perform constant, non-differential error simulations, where the α's—indicating the misclassification rate for source 1 and 2—are set so that $\alpha_1 = .35$ and $\alpha_2 = .2$. The replication of the bias estimates for β_0 and β_1 provided[2] in Cook et al. (2017) are presented for reference in A.1.[3]

Table A.2 presents the extension results of the simulation for estimates of β_0 and β_1 when $\alpha = .85$ and $\alpha_2 = .9$ (i.e., severe underreporting in both sources) in Experiment 1.[4] Briefly, we can see that for all models, except for Model 3, the bias, standard deviation, and standard error for the estimate of β_0 increase. For Model 3, the Hausman model with covariates, the bias, standard deviation, and standard error for Model 3 all decrease. However, the magnitude of the standard deviation and standard error are such that this estimate is still fairly imprecise. Interestingly, after introducing severe underreporting in both sources, the bias switches directions (relative to the Cook et al.'s original simulation results) for both

stages.

[2] I.e., the coefficient estimates for the constant term and covariate related to the occurrence of one's actual process or event of interest.

[3] Note that the values for Model 4 are slightly different from those reported in Cook et al. (2017).

[4] Alternatively, Table A.5 presents both the original simulation results and the adjusted simulation results side-by-side.

Table A.1: Replication of Simulation Results (Table 1, p231)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .35$, $\alpha_2 = .2$			
β_0 = -1	Bias	0.052	-0.007	-1.704	0.004	0.002
	STD	0.058	0.092	8.472	0.063	0.064
	SE	0.060	0.092	3.336	0.063	0.064
	MSE	0.006	0.009	74.604	0.004	0.004
	CP (%)	87.500	92.177	95.792	95.900	95.700
$\beta_1 = 1$	Bias	0.054	-0.028	-1.388	-0.007	-0.005
	STD	0.066	0.111	9.308	0.075	0.078
	SE	0.067	0.110	3.028	0.077	0.079
	MSE	0.007	0.013	88.478	0.006	0.006
	CP (%)	85.700	92.979	93.086	95.500	95.300

Table A.2: Adjusted Simulation Results (Table 1, p231)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .85$, $\alpha_2 = .9$			
β_0 = -1	Bias	0.798	-1.698	-0.752	-0.027	-0.364
	STD	0.082	46.201	7.957	0.303	1.123
	SE	0.086	108.407	2.510	0.274	0.743
	MSE	0.644	2135.244	63.809	0.093	1.392
	CP(%)	0.000	80.300	30.589	92.593	88.608
$\beta_1 = 1$	Bias	0.469	-11.854	-2.289	-0.117	-0.157
	STD	0.073	322.358	15.710	0.514	1.203
	SE	0.079	1172.196	4.596	0.318	0.584
	MSE	0.225	103951.402	251.805	0.277	1.470
	CP(%)	0.000	82.400	44.614	93.694	91.552

of Cook et al.'s multi-source models (Models 4 and 5) moving from a slight positive bias to a larger negative bias in these estimates. Under our first extension, the mean squared error increases for all models except for Model 3. However, the magnitude of the change differs across models, with Model 2 notably increasing from .009 to 2135.244. The coverage probabilities also decrease for every model, as expected, though again with wildly differing magnitudes. For example, the naive probit (Model 1) sees its coverage probabilities approach 0 percent, whereas the Hausman with covariates model (i.e., Model 3) decreases in coverage probabilities to around 30 percent.

For β_1, the changes under extension 1 are similar in direction, but not magnitude, across the models: the bias, standard deviation, standard error, and mean squared error each *increase*. The coverage probabilities again decrease for all models. Overall, Model 4—Cook et al.'s multi-source model with constant probabilities—seems to perform the best in estimating both β_0 and β_1 under circumstances of severe underreporting in two sources.

Table A.3 presents the replicated marginal effects provided in Cook et al. (2017). Table A.4 then presents the marginal effects, $\partial Y/\partial X$, for Experiment 1 when $\alpha_1 = .85$ and $\alpha_2 = .9$ (i.e., for extension 1).[5] Intuitively, we see that for all models, except for Model 2, the bias of the marginal effect increases under circumstances of severe underreporting in two binary sources. The bias decreases for Model 2, but also switches from a positive to negative bias, with both the standard deviation and mean squared error increasing. For both of Cook et al.'s multi-source models, the bias increases, along with the standard deviation and mean squared error, with the bias for Model 5 switching from positive to negative.

Table A.3: Replication of Marginal Effects in Simulation Studies (Table 2, p232)

Parameter		(1) Naive Probit	(2) Hausman Const Pr	(3) Hausman w/ Cov	(4) Multi-source Const Pr	(5) Multi-source w/ Cov
				Experiment 1: $\alpha_1 = .35$, $\alpha_2 = .2$		
$\partial Y/\partial X$	Bias	-0.030	0.011	-0.038	0.002	0.001
	SE	0.023	0.049	0.109	0.027	0.028
	MSE	0.001	0.002	0.013	0.001	0.001

Tables A.5-A.7 provide the results of the two simulations (i.e., Cook et al. (2017)'s original Experiment 1 and our first extension) side by side. The original results ($\alpha_1 = .35$, $\alpha_2 = .2$) are presented in the white columns; the adjusted simulation results ($\alpha_1 = .85$, $\alpha_2 = .9$) are

[5] Alternatively, Table A.6 presents both the original simulation results and the adjusted simulation marginal effects side-by-side.

Table A.4: Adjusted Marginal Effects in Simulation Studies (Table 2, p232)

Parameter		(1) Naive Probit	(2) Hausman Const Pr	(3) Hausman w/ Cov	(4) Multi-source Const Pr	(5) Multi-source w/ Cov
				Experiment 1: $\alpha_1 = .85$, $\alpha_2 = .9$		
$\partial Y/\partial X$	Bias	-0.275	-0.005	-0.187	0.020	-0.040
	STD	0.011	0.172	0.145	0.126	0.172
	MSE	0.076	0.030	0.056	0.016	0.031

presented in the gray columns. Figures A.1 and A.2 present the β's and marginal effects simulation results, with the original results in black and the adjusted in blue. The dots are the bias for each model, with the bars representing the range covered by one standard deviation below and one standard deviation above. For purposes of clarity, the graphs are scaled so that the bars for the adjusted simulation of Model 2[6] are outside the range.[7] As noted earlier, these Tables and Figures largely confirm and re-present the patterns discussed above.

Table A.5: Comparison of Simulation Results (Table 1, p231)

		Naive Probit		Hausman Const Pr		Hausman w/ Cov		Multi-source Const Pr		Multi-source w/ Cov	
		(1)	(1a)	(2)	(2a)	(3)	(3a)	(4)	(4a)	(5)	(5a)
β_0 = -1	Bias	0.052	0.798	-0.007	-1.698	-1.704	-0.752	0.004	-0.027	0.002	-0.364
	STD	0.058	0.082	0.092	46.201	8.472	7.957	0.063	0.303	0.064	1.123
	SE	0.060	0.086	0.092	108.407	3.336	2.510	0.063	0.274	0.064	0.743
	MSE	0.006	0.644	0.009	2135.244	74.604	63.809	0.004	0.093	0.004	1.392
	CP(%)	87.500	0.000	92.177	80.300	95.792	30.589	95.900	92.593	95.700	88.608
$\beta_1 = 1$	Bias	0.054	0.469	-0.028	-11.854	-1.388	-2.289	-0.007	-0.117	-0.005	-0.157
	STD	0.066	0.073	0.111	322.358	9.308	15.710	0.075	0.514	0.078	1.203
	SE	0.067	0.079	0.110	1172.196	3.028	4.596	0.077	0.318	0.079	0.584
	MSE	0.007	0.225	0.013	103951.402	88.478	251.805	0.006	0.277	0.006	1.470
	CP(%)	85.700	0.000	92.979	82.400	93.086	44.614	95.500	93.694	95.300	91.552

White columns: Experiment 1 ($\alpha_1 = .35$, $\alpha_2 = .2$). Gray columns: Experiment 1 ($\alpha_1 = .85$, $\alpha_2 = .9$).

[6] For β_0 they range from -47.899 to 44.503; for β_1 they range from -334.212 to 310.504.

[7] Graphs at natural scale are provided in Figure A.3.

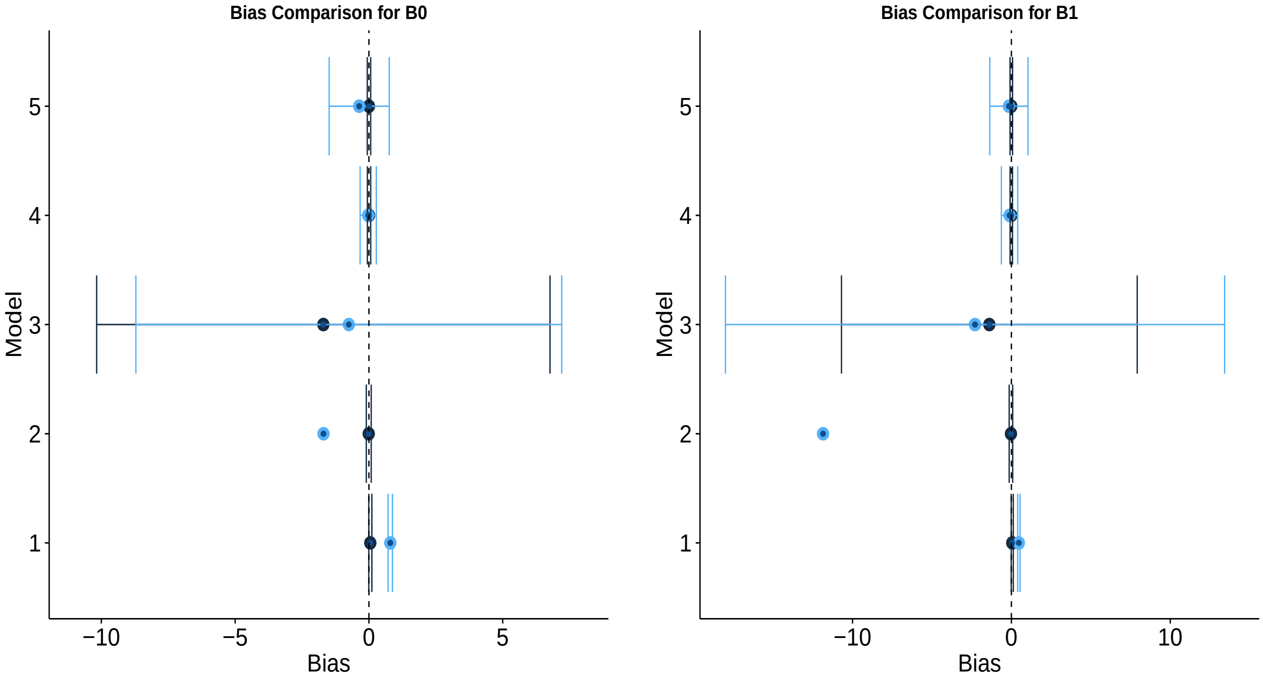

Figure A.1: Comparison of Bias for Original Simulation (Black) with Adjusted (Blue)

Table A.6: Comparison of Marginal Effects in Simulation Studies (Table 2, p232)

		Naive Probit		Hausman Const Pr		Hausman w/ Cov		Multi-source Const Pr		Multi-source w/ Cov	
		(1)	(1a)	(2)	(2a)	(3)	(3a)	(4)	(4a)	(5)	(5a)
$\partial Y/\partial X$	Bias	-0.030	-0.275	0.011	-0.005	-0.038	-0.187	0.002	0.020	0.001	-0.040
	STD	0.023	0.011	0.049	0.172	0.109	0.145	0.027	0.126	0.028	0.172
	MSE	0.001	0.076	0.002	0.030	0.013	0.056	0.001	0.016	0.001	0.031

White columns: Experiment 1 ($\alpha_1 = .35$, $\alpha_2 = .2$). Gray columns: Experiment 1 ($\alpha_1 = .85$, $\alpha_2 = .9$).

Table A.7: Comparison of Simulation Results (Table 1, p231)

		Naive Probit		Hausman Const Pr		Hausman w/ Cov		Multi-source Const Pr		Multi-source w/ Cov	
		(1)	(1a)	(2)	(2a)	(3)	(3a)	(4)	(4a)	(5)	(5a)
β_0 = -1	Bias	0.052	0.798	-0.007	-1.698	-1.704	-0.752	0.004	-0.027	0.002	-0.364
	STD	0.058	0.082	0.092	46.201	8.472	7.957	0.063	0.303	0.064	1.123
	SE	0.060	0.086	0.092	108.407	3.336	2.510	0.063	0.274	0.064	0.743
	MSE	0.006	0.644	0.009	2135.244	74.604	63.809	0.004	0.093	0.004	1.392
	CP(%)	87.500	0.000	92.177	80.300	95.792	30.589	95.900	92.593	95.700	88.608
	MAE	0.063	0.798	0.072	2.130	1.746	1.942	0.050	0.229	0.051	0.615
$\beta_1 = 1$	Bias	0.054	0.469	-0.028	-11.854	-1.388	-2.289	-0.007	-0.117	-0.005	-0.157
	STD	0.066	0.073	0.111	322.358	9.308	15.710	0.075	0.514	0.078	1.203
	SE	0.067	0.079	0.110	1172.196	3.028	4.596	0.077	0.318	0.079	0.584
	MSE	0.007	0.225	0.013	103951.402	88.478	251.805	0.006	0.277	0.006	1.470
	CP(%)	85.700	0.000	92.979	82.400	93.086	44.614	95.500	93.694	95.300	91.552
	MAE	0.070	0.469	0.086	12.095	1.563	2.839	0.061	0.270	0.063	0.533

White columns: Experiment 1 ($\alpha_1 = .35$, $\alpha_2 = .2$). Gray columns: Experiment 1 ($\alpha_1 = .85$, $\alpha_2 = .9$).

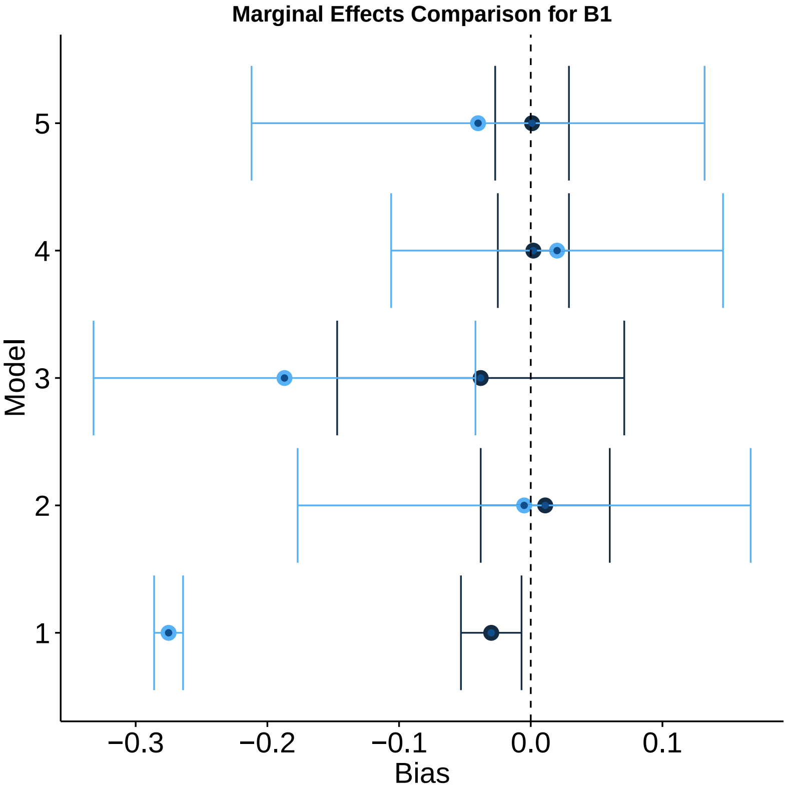

Figure A.2: Comparison of Marginal Effects Original Simulation (Black) with Adjusted (Blue)

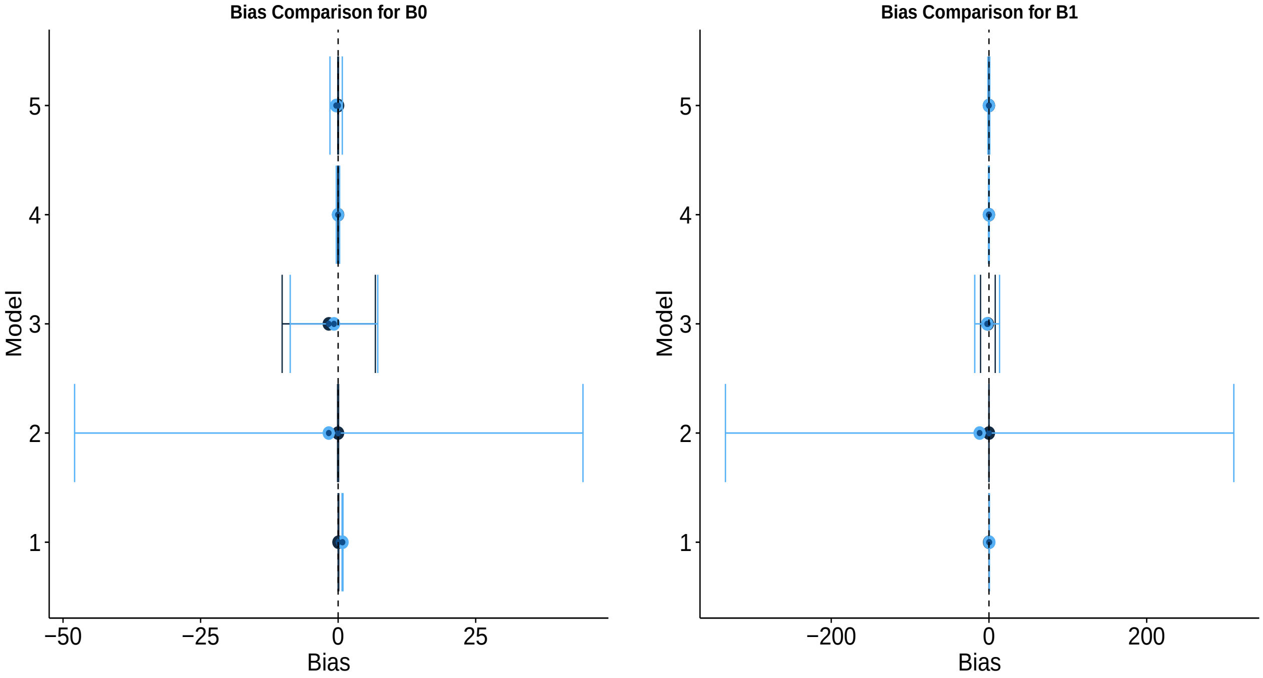

Figure A.3: Comparison of Bias for Original Simulation (Black) with Adjusted (Blue) with Model 2 Standard Deviation

Extension 2

Our second extension reconsiders Cook et al.'s Experiment 1 under circumstances where the researcher encounters one highly accurate reporting source and one extremely inaccurate reporting source. For reference, Table A.8[8] provides the original results of the simulation performed in Cook et al. (2017), where the misclassification rates for sources 1 and 2 are set to $\alpha_1 = .35$ and $\alpha_2 = .2$, respectively. As mentioned above, the goal of our second extension is to investigate the performance of the Cook et al. (2017) estimators when the rates of misclassification among two distinct reporting sources are substantially different from one another. Correspondingly, Table A.9 presents the results of this extension, specifically for estimates of β_0 and β_1 when $\alpha = .1$ and $\alpha_2 = .9$, while maintaining all other quantities to the levels assigned by Cook et al. under their Experiment 1.[9]

Table A.8: Replication of Simulation Results (Table 1, p231)

		(1)	(2)	(3)	(4)[10]	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
				Experiment 1: $\alpha_1 = .35$, $\alpha_2 = .2$		
β_0 = -1	Bias	0.052	-0.007	-1.704	0.004	0.002
	STD	0.058	0.092	8.472	0.063	0.064
	SE	0.060	0.092	3.336	0.063	0.064
	MSE	0.006	0.009	74.604	0.004	0.004
	MAE	0.063	0.072	1.746	0.050	0.051
	CP (%)	87.500	92.177	95.792	95.900	95.700
$\beta_1 = 1$	Bias	0.054	-0.028	-1.388	-0.007	-0.005
	STD	0.066	0.111	9.308	0.075	0.078
	SE	0.067	0.110	3.028	0.077	0.079
	MSE	0.007	0.013	88.478	0.006	0.006
	MAE	0.070	0.086	1.563	0.061	0.063
	CP (%)	85.700	92.979	93.086	95.500	95.300

Briefly, we see in our second extension that the bias for β_0 and β_1 increases slightly for

[8] Note: we have also added values for the Mean Average Error (MAE) to this Table.

[9] Alternatively, Table A.12 presents both the original simulation results and the adjusted simulation results side-by-side. Additionally, Table A.14 provides the results for the first extension where $\alpha_1 = .85$ and $\alpha_2 = .9$ for reference.

Table A.9: Results of Extension (Table 1, p231)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .1$, $\alpha_2 = .9$			
β_0 = -1	Bias	0.066	-0.001	-1.519	0.006	-0.029
	STD	0.058	0.097	7.570	0.073	0.111
	SE	0.060	0.097	2.836	0.074	0.099
	MSE	0.008	0.009	59.549	0.005	0.013
	MAE	0.074	0.076	1.573	0.057	0.080
	CP(%)	83.100	90.700	93.681	94.100	97.202
$\beta_1 = 1$	Bias	0.068	-0.025	-1.139	-0.008	0.026
	STD	0.065	0.120	8.412	0.093	0.124
	SE	0.067	0.114	2.484	0.088	0.101
	MSE	0.009	0.015	71.993	0.009	0.016
	MAE	0.079	0.092	1.339	0.073	0.096
	CP(%)	81.600	92.600	91.775	94.200	90.754

Model 1 and Model 4, while it decreases for Models 2 and 3.[11] The bias in Model 5 both increases and changes direction, moving from positive to negative for β_0 and from negative to positive for β_1. However, in all models, except for Model 3, the uncertainty increases, though in some cases at in the ten-thousandth decimal place. Similarly, the Mean Squared Error (MSE) and Mean Average Error (MAE) increase for all models except Model 3. Interestingly, the coverage probability increases for Model 5 in the case of β_0, bit decreases for all other models for both β_0 and β_1.

Table A.10 presents the replicated marginal effects provided in Cook et al. (2017). Table A.11 presents the marginal effects, $\partial Y/\partial X$, of Experiment 1 when $\alpha_1 = .1$ and $\alpha_2 = .9$.[12] The bias increases for Model 1, Model 3, and Model 5, with the bias changing from positive to negative for Model 5. There is a slight decrease in the bias for Model 2, and the change for Model 4 is in the ten-thousandth decimal place. We also find that —in circumstances of one fairly accurate reporting source and one very poor reporting source—the uncertainty for

[11] That is, relative to Cook et al.'s original results.

[12] Alternatively, Table A.13 presents both the original simulation results and the adjusted simulation marginal effects side-by-side.

all models increases, with both the standard deviation and MSE increasing from the original simulation.

Table A.10: Replication of Marginal Effects in Simulation Studies (Table 2, p232)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .35$, $\alpha_2 = .2$			
$\partial Y/\partial X$	Bias	-0.030	0.011	-0.038	0.002	0.001
	SE	0.023	0.049	0.109	0.027	0.028
	MSE	0.001	0.002	0.013	0.001	0.001

Table A.11: Marginal Effects in Extension of Simulation Studies (Table 2, p232)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .1$, $\alpha_2 = .9$			
$\partial Y/\partial X$	Bias	-0.038	0.008	-0.044	0.002	-0.007
	STD	0.023	0.053	0.115	0.038	0.044
	MSE	0.002	0.003	0.015	0.001	0.002

To compare all the models, experiments, and extensions considered here, Figures A.4 and A.5 present the β's and marginal effects results for the original simulation where $\alpha_1 = .35, \alpha_2 = .2$ (black), the first extension where $\alpha_1 = .85, \alpha_2 = .9$ (yellow), and the $\alpha_1 = .1, \alpha_2 = .9$ (blue). The dots are the bias for each model, with the bars representing the range covered by one standard deviation below and one standard deviation above. The x-axis in for the left panel in A.4 ranges from -11 to 8, while the right panel ranges from -18 to 14 so that the differences between estimates can be seen more clearly.[13] However, this means that the standard deviations for Model 2 in extension 1 are not reported in the figure as they are substantially outside the range.[14] Compared to the first extension, our second extension maps closely on to Cook et al.'s original simulation results, both with regards to bias and

[13] Graphs with the full range are provided in Figure A.6.

[14] For β_0 they range from -47.899 to 44.503; for β_1 they range from -334.212 to 310.504.

uncertainty. The only case where the first extension outperforms both the original and the second extension is β_0 in Model 3. In all cases, the second extension has less uncertainty around the estimates.

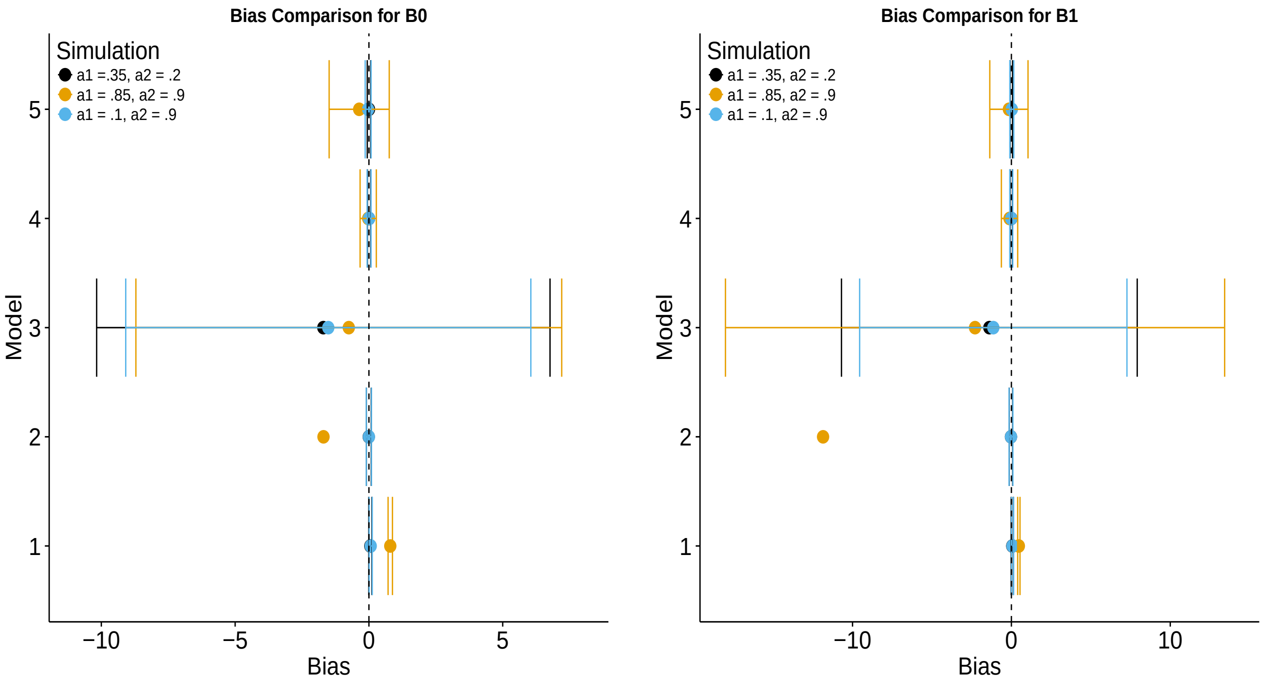

Figure A.4: Comparison of Bias for Original Simulation (Black) with Extension 1 (Yellow) and Extension 2 (Blue) with reduced x-axis range

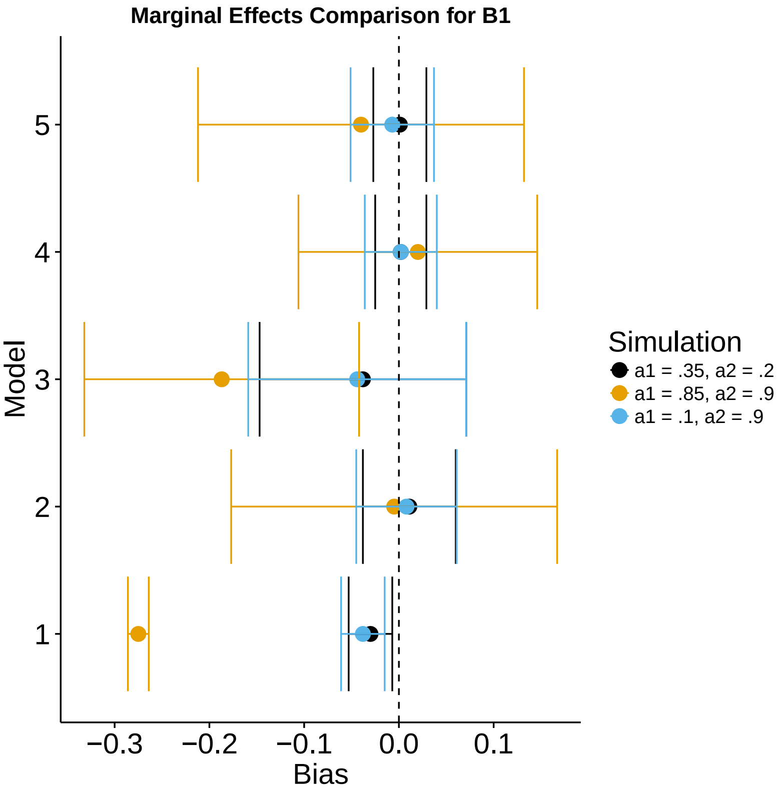

Figure A.5: Comparison of Marginal Effects Original Simulation (Black) with Extension 1 (Yellow) and Extension 2 (Blue)

For further reference, Table A.12 and A.13 provide the results of the two simulations side by side. The original results ($\alpha_1 = .35$, $\alpha_2 = .2$) are presented in the white columns; extension 2's adjusted simulation results ($\alpha_1 = .1$, $\alpha_2 = .9$) are presented in the gray columns.

Table A.12: Comparison of Simulation Results (Table 1, p231)

		Naive		Hausman		Hausman		Multi-source		Multi-source	
		Probit		Const Pr		w/ Cov		Const Pr		w/ Cov	
		(1)	(1a)	(2)	(2a)	(3)	(3a)	(4)	(4a)	(5)	(5a)
β_0 = -1	Bias	0.052	0.066	-0.007	-0.001	-1.704	-1.519	0.004	0.006	0.002	-0.029
	STD	0.058	0.058	0.092	0.097	8.472	7.57	0.063	0.073	0.064	0.111
	SE	0.060	0.06	0.092	0.097	3.336	2.836	0.063	0.074	0.064	0.099
	MSE	0.006	0.008	0.009	0.009	74.604	59.549	0.004	0.005	0.004	0.013
	MAE	0.063	0.074	0.072	0.076	1.746	1.573	0.050	0.057	0.051	0.08
	CP(%)	87.500	83.1	92.177	90.7	95.792	93.681	95.900	94.1	95.700	97.202
$\beta_1 = 1$	Bias	0.054	0.068	-0.028	-0.025	-1.388	-1.139	-0.007	-0.008	-0.005	0.026
	STD	0.066	0.065	0.111	0.12	9.308	8.412	0.075	0.093	0.078	0.124
	SE	0.067	0.067	0.110	0.114	3.028	2.484	0.077	0.088	0.079	0.101
	MSE	0.007	0.009	0.013	0.015	88.478	71.993	0.006	0.009	0.006	0.016
	MAE	0.070	0.079	0.086	0.092	1.563	1.339	0.061	0.073	0.063	0.096
	CP(%)	85.700	81.6	92.979	92.6	93.086	91.775	95.500	94.2	95.300	90.754

White columns: Experiment 1 ($\alpha_1 = .35$, $\alpha_2 = .2$). Gray columns: Experiment 1 ($\alpha_1 = .1$, $\alpha_2 = .9$).

Table A.13: Comparison of Marginal Effects in Simulation Studies (Table 2, p232)

		Naive		Hausman		Hausman		Multi-source		Multi-source	
		Probit		Const Pr		w/ Cov		Const Pr		w/ Cov	
		(1)	(1a)	(2)	(2a)	(3)	(3a)	(4)	(4a)	(5)	(5a)
$\partial Y/\partial X$	Bias	-0.030	-0.038	0.011	0.008	-0.038	-0.044	0.002	0.002	0.001	-0.007
	STD	0.023	0.023	0.049	0.053	0.109	0.115	0.027	0.038	0.028	0.044
	MSE	0.001	0.002	0.002	0.003	0.013	0.015	0.001	0.001	0.001	0.002

White columns: Experiment 1 ($\alpha_1 = .35$, $\alpha_2 = .2$). Gray columns: Experiment 1 ($\alpha_1 = .1$, $\alpha_2 = .9$).

Figure A.6 provides the estimates for the bias (dots) with the corresponding one standard deviation below and one standard deviation above (bars). The x-axis is expanded so that the standard deviations for Model 2 are plotted on the graph.

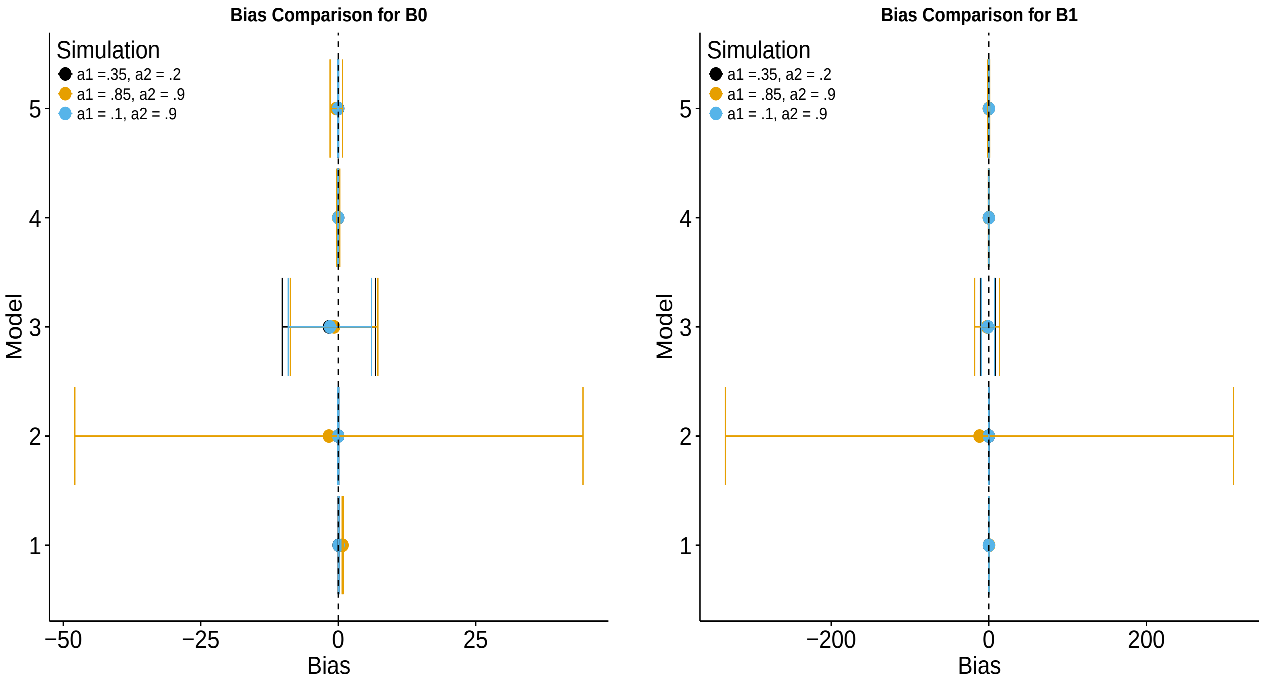

Figure A.6: Comparison of Bias for Original Simulation (Black) with Extension 1 (Yellow) and Extension 2 (Blue) with the full range of the x-axis

Table A.14 provides the results from the first extension where $\alpha_1 = .85, \alpha_2 = .9$. As reflected in the graphs, the results differ more from the original results than in the case of the second simulation where $\alpha_1 = .1, \alpha_2 = .9$.

In summary, our extensions of Cook et al.'s Monte Carlo experiments provide several salient insights. First, we find that in situations of extreme underreporting in two independent sources and constant misclassification processes, the Cook et al. multi-source models typically outperform plausible alternative models such as naive probits and Hausman estimators. However, in these circumstances, the analyst may be best off in using Cook et al.'s multi-source models with constant probabilities, as opposed to the multi-source models that include covariates in the misclassification stage, as the former multi-source specification performs best in estimating both β_0 and β_1 under circumstances of severe underreporting in

Table A.14: Results of Extension 1 (Table 1, p232)

		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-source	Multi-source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Experiment 1: $\alpha_1 = .85$, $\alpha_2 = .9$			
β_0 = -1	Bias	0.798	-1.698	-0.752	-0.027	-0.364
	STD	0.082	46.201	7.957	0.303	1.123
	SE	0.086	108.407	2.510	0.274	0.743
	MSE	0.644	2135.244	63.809	0.093	1.392
	MAE	0.798	2.130	1.942	0.229	0.615
	CP(%)	0.000	80.300	30.589	92.593	88.608
$\beta_1 = 1$	Bias	0.469	-11.854	-2.289	-0.117	-0.157
	STD	0.073	322.358	15.710	0.514	1.203
	SE	0.079	1172.196	4.596	0.318	0.584
	MSE	0.225	103951.402	251.805	0.277	1.470
	MAE	0.469	12.095	2.839	0.270	0.533
	CP(%)	0.000	82.400	44.614	93.694	91.552

both (i.e., two) sources. Second, and comparing our results across both extensions, we also intuitively find that having one good source and one bad source is preferable to having two bad sources, but that either scenario is less preferable to having two moderate misclassification sources. Nevertheless, in each extension, we conclude that the Cook et al. model still outperforms all binary comparison models with respect to bias. This implies that even with two bad sources (or with one good source and one bad source), the Cook et al. estimator is still preferable to analyzing one's binary (collapsed) data with single-source models.

Application 1: Repression in Africa

Recall that our first event data comparison considered the case of state repression in Africa. As we noted in our main Research Note, our analysis in this regard builds on that of Cook et al. (2017), who examine the prevalence of reporting bias within the context of monthly instances of state repression across African countries for the years 2012-2013. The authors do so with the aid of the Social Conflict Analysis Database (SCAD; Hendrix et al., 2012). SCAD is a human coded event dataset that records a wide range of political and social conflict under the "who did what to whom (and where/when)" relational event framework. It provides an ideal application of the misclassification methods developed by Cook et al. (2017) given its inclusion of an indicator variable recording whether (or not) each coded event was identified in (i) the Associated Press (AP) or (ii) Agence-France Presse (AFP); and also given past studies establishing the presence of reporting bias within the SCAD data (Hendrix and Salehyan, 2015). In light of these advantages, our main Note accordingly replicates Cook et al.'s study of state repression in Africa, both with the human-coded SCAD data originally used by the authors and when using a comparably disaggregated set of machine-coded state repression events derived from the World-Integrated Crisis Early Warning System Dataset (ICEWS; Boschee et al., 2016).

In this section of our Supplemental Appendix, we seek to provide a more detailed overview of the data, aggregation decisions, and analyses used within this application. We begin first by reviewing the machine coded event data that are used in the current application. The ICEWS dataset is a fully machine-coded, CAMEO-based (Schrodt, Gerner and Yilmaz, 2009), event dataset that draws upon approximately 300 electronically available news sources to code relational events at a global scale for the years 1995-Present. Similar to SCAD, ICEWS includes an indicator variable that records the specific newswire or news agency that was used to code a given event. This feature allows us to recover and retain only those ICEWS "state repression" events for African countries that were coded from AFP and AP. When combined with the SCAD data described above, we are thus able to make the controlled comparisons of human and machine coded event data that are reported in our main Note, for the same news sources, locations, time-frames, source/target-actors, and event types. After accounting for underreporting issues within our human and machine coding event data, these comparisons provide us with a sense of the relative quality of modern human and machine coded event data for political violence within Africa. Below we describe our Africa repression data — and corresponding data aggregation tasks— in further detail, before proceeding to a full summary of our multi-source models and related validation comparisons.

Data Formatting

In this subsection, we first briefly describe the formatted SCAD data used by Cook et al. (2017). We then discuss how we formatted the machine-coded ICEWS data to match Cook et al.'s cases of state repression in Africa. The SCAD data used by Cook et al. (2017) record events for 47 African countries during the years 2012-2013. These event data are aggregated to the country-month level by Cook et al. (2017), and are then subset to include only (i) those events initiated by government or pro-government actors and (ii) event types corresponding to lethal or non-lethal repression. Within the SCAD codebook, repression is defined as "[d]istinct violent event waged primarily by government authorities, or by groups acting in explicit support of government authority, targeting individual, or 'collective individual,' members of an alleged opposition group or movement" (Hendrix and Salehyan, 2012, 3). Examples provided by this codebook of nonlethal repression include tear-gas and arrests, whereas lethal repression must include casualties. Our detailed examination of the nonlethal repression events included within the SCAD dataset suggests that these events correspond to material instances of repression (e.g., arrests, the breaking up of protests, etc.) rather than verbal threats of repression, although some rare instances of nonlethal repression within the SCAD codebook could be interpreted as a show or threat of force (e.g., deploying tanks).

Returning to Cook et al. (2017), we note that the authors then dichotomize their resulting country-month SCAD repression event counts to create a primary repression dependent variable, set equal to one for African country-months that experienced at least one state repression event, and zero otherwise. Cook et al. (2017) then create two separate versions of this dichotomous country-month variable for use in their misclassification-model. The first binary dependent variable equals one for only the African repression cases that SCAD derived from AP news stories. The second binary dependent variable is set equal to one for only those repression cases that SCAD derived from AFP-published stories. These are the SCAD-based repression measures that we use in our multi-source misclassification models below.

To do so, we downloaded Cook et al.'s formatted SCAD replication data from Harvard's Dataverse. We then sought to format the World ICEWS data to correspond as closely as possible to Cook et al.'s African state repression event indicators. Because the default ICEWS actor and action designations do not perfectly match those used by SCAD, we had to make a number of aggregation decisions when formatting our ICEWS data. In doing so, we made every effort to be as comprehensive as possible in retaining all relevant source and target actors for comparison, while also not including any possibly questionable ICEWS-actor designations for our SCAD comparisons. We first retained only those ICEWS events arising from government actors within African countries for the years 2012-2013.[15] We then retained only those ICEWS events that contained a domestic social actor as a target.[16] With these source-target pairings in hand, we subset our remaining ICEWS event data to only include repression conflict based CAMEO-based ICEWS events. To do so, we included all material (i.e., non-verbal) conflict CAMEO-category coded ICEWS events aside from "14: PROTEST," which is a decidedly citizen-directed form of material conflict, rather than a realistic category for material acts of violence used by governments for the purposes of repression. The final set of two-digit CAMEO categories included in our ICEWS measure of

[15] Including source actors designated as "Military," "Government," or "Police;" or those ambiguously assigned to have a country as a source actor, with no additional details of the nature of that source actor.

[16] Including target actors designated as "General Population," "Civilian," "Social," "Protestors," "Mobs," or "Popular Opposition."

repression are:[17]

15: EXHIBIT FORCE POSTURE
17: COERCE
18: ASSAULT
19: FIGHT
20: USE UNCONVENTIONAL MASS VIOLENCE

These retained ICEWS events were then divided into two separate (source-specific) event datasets. The first ICEWS datset corresponds to only those ICEWS events coded from the AP. The second ICEWS event dataset corresponds to events coded from AFP newswires. Importantly, our subsetting of ICEWS into separate event datasets for AP and AFP derived events retains only a small number of all African state repression events coded by ICEWS. Specifically, for our two years of interest, ICEWS drew from a total of 136 distinct news sources in recording African state repression events. Of the 46,485 African state repression events coded by ICEWS (based upon our definition of state repression) from these 136 sources for 2012-2013, only 6,726 (14.5%) were coded from AP or AFP sources. As such, substantial underreporting in our ICEWS data exists *by design*, and is intended to maximize comparability of the ICEWS data with the formatted SCAD data used by Cook et al. (2017); as well as to facilitate the usefulness of the multi-source models described earlier for these purposes. Any findings of underreporting in our final ICEWS data should therefore not be interpreted as indicative of comparable levels of underreporting within the complete ICEWS dataset.

With the above points in mind, our two news source-specific ICEWS event datasets were aggregated to the country-month level for all African countries.[18] Following this, we dichotomized our AP and AFP newswire sources to correspond to comparable country-month indicators of state repression in Africa to the SCAD indicators used by Cook et al. (2017). We then merged the formatted ICEWS and SCAD African repression data together for

[17] Note that we also included the three digit CAMEO subcategories to each two-digit category as well.

[18] Note that although duplicate events are a concern in ICEWS, this is not an issue for our aggregated repression data, given in this case we simply code whether or not an African country-year experienced at least one repression event.

Table A.15: Dichotomized Country-Month Summary Statistics (2012-2013)

	Mean	Stdev	Min	Max	Total Conflict Cases
SCAD (AP)	0.040	0.197	0	1	44
SCAD (AFP)	0.070	0.255	0	1	76
ICEWS (AP)	0.137	0.343	0	1	149
ICEWS (AFP)	0.298	0.457	0	1	325

Note: $N = 1,092$

analysis. Below we briefly describe a series of univariate and bivariate descriptive statistics for our merged ICEWS and SCAD data, before presenting our model-based validation tasks.

Dichotomized Repression Measure Comparisons

For summary purposes, this section compares the dichotomized African repression indicators discussed above via descriptive statistics and bivariate comparisons. These comparisons examine the specific 2012-2013 African country-month sample used by Cook et al. (2017). Cook et al. (2017) have 1,092 cases in their final analysis, which they mention corresponds to 47 African countries. Upon closer examination, it appears that the final Cook et al. (2017) analysis sample actually corresponds to 46 African countries (after listwise deletion), with each generally observed for all 24 months. Table A.15 reports univariate summary statistics for the state repression indicators contained in this African sample. We generally find that ICEWS records substantially more African country-months as experiencing at least one repression event during our period of analysis. A portion of these discrepancies may be due to the more fine grained categories of repression included in the CAMEO coding scheme, relative to SCAD. Nevertheless, given that both datasets purport to capture both lethal and nonlethal material repression against domestic actors based on the event action categories chosen, it would appear that the machine-coded ICEWS data outperform SCAD in their net coverage of African repression events.

The confusion matrices in Table A.16 allow us to directly compare the (overlapping) event coverage across news sources, both (i) within the SCAD or ICEWS data (Tables A.16a-A.16b) and (ii) for each news source across the SCAD/ICEWS datasets (Tables A.16c-A.16d). Starting with Tables A.16a-A.16b, we find that SCAD contains 23 country-month instances of state repression that were uniquely identified AP sources, 55 state-repression country-months that were uniquely identified as such by the AFP, and 21 instances where AP and AFP jointly recorded a state repression country-month in SCAD. By comparison, our ICEWS data contain 43 state repression country-months uniquely that were uniquely derived from the AP, 219 state repression country-months uniquely derived from AFP, and 106 state repression country-months that were jointly recorded in both AP and AFP. The level of AP-AFP overlap within our ICEWS data's country-month repression indicators is therefore higher (in absolute terms), when compared to the SCAD data. Our news source specific comparisons (i.e., our comparisons of the same news sources across event datasets) in Tables A.16c-A.16d reveal that the vast majority of country-months coded as exhibiting state repression by our AP-specific, or AFP-specific, sources are distinct within SCAD and ICEWS data. That is, ICEWS and SCAD appear to be largely identifying different sets of repressive country-months within Africa for the years 2012-2013. This suggests that there is substantial underreporting within each source considered here. As demonstrated in extension 1 to our Monte Carlo experiments, severe underreporting in both sources may correspond to higher levels of bias and uncertainty in one's multi-source estimates than otherwise, yet these estimates will still be generally preferable to those obtained from non-multi-source estimators.

Multi-source Model Comparisons

As discussed in our main Note, our full model-based comparisons of the SCAD and ICEWS African repression data proceed in several steps. We first replicate Cook et al.'s SCAD application using their proposed multi-source estimator. We then repeat this exercise when using the ICEWS data in place of SCAD. In each case, we follow Cook et al. (2017) by first estimating a set of *multi-source constant* specifications. These specifications include only constant terms within the misclassification stages of the Cook et al. multi-source estimator, and include the following covariates within the repression stage of the model:

Table A.16: Dichotomized Country-Month Confusion Matrices for State Repression (2012-2013)

(a) SCAD (AP) Vs. SCAD (AFP)

		AFP		
		0	1	Total
AP	0	993	55	1,048
	1	23	21	44
	Total	1,016	76	1,092

Pearson $\chi^2 = 117.674$, $P < 0.001$

(b) ICEWS (AP) Vs. ICEWS (AFP)

		AFP		
		0	1	Total
AP	0	724	219	943
	1	43	106	149
	Total	767	325	1,092

Pearson $\chi^2 = 141.327$, $P < 0.001$

(c) SCAD (AP) Vs. ICEWS (AP)

		ICEWS		
		0	1	Total
SCAD	0	924	124	1048
	1	19	25	44
	Total	943	149	1,092

Pearson $\chi^2 = 72.526$, $P < 0.001$

(d) SCAD (AFP) Vs. ICEWS (AFP)

		ICEWS		
		0	1	Total
AFP	0	745	271	1,016
	1	22	54	76
	Total	767	325	1,092

Pearson $\chi^2 = 66.622$, $P < 0.001$

GDP per capita$_{t-1}$, *Population*$_{t-1}$, and *Democracy*$_{t-1}$.[19] For each dependent variable (i.e., SCAD and ICEWS), we then estimate *multi-source with covariates* specifications that include *GDP per capita*$_{t-1}$, *Population*$_{t-1}$, and *Democracy*$_{t-1}$ in both the repression and misclassification stages of the Cook et al. multi-source estimators, while also adding AFP Reports and AP Reports[20] to the relevant misclassification stages of these estimators. These models appear in Table A.17-A.18.

Beginning with Table A.17, we find that the estimated effects of *GDP per capita*$_{t-1}$, *Population*$_{t-1}$, *Democracy*$_{t-1}$ on repression are remarkably similar across our SCAD and ICEWS specifications. The most notable difference across our SCAD and ICEWS specifications is that of *GDP per capita*$_{t-1}$, which is positive and not statistically significant in the SCAD *multi-source constant* specification, but positive and statistically significant ($p < 0.01$) in the ICEWS *multi-source constant* specification; implying that more developed

[19] The operationalizations of each variable are described in Cook et al. (2017), who expect *GDP per capita*$_{t-1}$ and *Democracy*$_{t-1}$ to each be negatively related to repression, but *Population*$_{t-1}$ to be positively associated with repression.

[20] These measures were collected by Cook et al. (2017), and report the number of non-conflict AFP and AP news reports for each country under analysis.

African countries are more likely to exhibit monthly repression.[21] However, in turning to the SCAD and ICEWS *multi-source with covariates* specifications in Table A.17, we now more find that *GDP per capita*$_{t-1}$ is negative and statistically significant ($p < 0.01$) in each case. This intuitively suggests that development is associated with less monthly repression, thereby underscoring the value-added of Cook et al.'s *multi-source with covariates* model over its *multi-source constant* counterpart. Note however that, given the likelihood of high underreporting among all sources examined here, and our Monte Carlo analyses of such circumstances, our multi-source estimates will possibly be fairly imprecise.

For *Population*$_{t-1}$, we find a positive and statistically significant effect ($p < 0.01$) within both SCAD models, and in both ICEWS models, in Table A.17. This implies that more populous countries are more likely to exhibit one or more repression events in any given month, even after correcting for reporting bias issues. This consistency in estimated effects—alongside those for *GDP per capita*$_{t-1}$ in the *multi-source with covariates* specification above—underscores the comparability of our estimates repression-determinants across both human- and machine-coded event data, when aggregated to the country-month level. Our findings for *Democracy*$_{t-1}$ reinforce these conclusions, as the coefficient estimates for this variable are negative and statistically significant ($p < 0.01$) in each SCAD and ICEWS specification in Table A.17. Intuitively this finding suggests that monthly instances of repression are significantly less likely in democracies. Altogether, our Table A.17 findings hence strongly indicate that using machine coded event data in place of human coded data for reporting-bias adjusted analyses of African repression yields comparable theoretical findings—especially in the *multi-source with covariates* specification context.

Table A.18 offers additional evidence for consistency in estimates derived from machine- and human coded event data. This table reports the AP- and AFP-misclassification equation estimates for the models reported in Table A.17. Again the most noticeable differences in

[21] For reference, we note that this variable was positive but not statistically significant in Cook et al.'s naïve probit specification, and was found to be positive and statistically significant within expanded analyses of SCAD-based repression in Africa (Hendrix and Salehyan, 2016).

Table A.17: Models of Repression in Africa 2012-2013

	SCAD	ICEWS	SCAD	ICEWS
	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)
	Multi-Source	Multi-Source	Multi-Source	Multi-Source
	Constant Pr	Constant Pr	W/ Cov	W/Cov
$GDPpc_{t-1}$	0.021	0.257	-0.292	-0.534
	(0.072)	(0.048)	(0.145)	(0.185)
Pop_{t-1}	0.458	0.500	0.330	0.859
	(0.063)	(0.043)	(0.095)	(0.121)
$Demo_{t-1}$	-0.756	-0.385	-0.819	-1.488
	(0.172)	(0.105)	(0.315)	(0.405)
Constant	-8.562	-9.904	-3.857	-8.159
	(1.160)	(0.837)	(2.063)	(1.846)

Note: $N = 1,092$. Values in parentheses are standard errors.

our SCAD- and ICEWS-based estimates arise in the case of $GDP\ per\ capita_{t-1}$. Looking specifically at the *multi-source with covariates* specifications, we find that more developed countries are significantly ($p < 0.10$) less likely to exhibit reporting bias in the AP equation of the SCAD specification, but that $GDP\ per\ capita_{t-1}$ is not statistically significant in the AFP equation of the SCAD specification. By comparison, $GDP\ per\ capita_{t-1}$ is negative and statistically significant in both the AP and AFP equations of the ICEWS-based *multi-source with covariates* misclassification stage. The coefficient estimate for $Population_{t-1}$ is consistently negative across the SCAD and ICEWS misclassification equations (implying that more populous countries are less likely to exhibit reporting bias), though it is not statistically significant within the SCAD AFP equation. $Democracy_{t-1}$ is not statistically significant in any of the SCAD or ICEWS misclassification equations. Finally, in all cases, we find that AP Reports and AFP Reports are consistently negative and statistically significant ($p < 0.01$) predictors of reporting bias. For a given African country-month, and no matter whether one examines human- or machine-coded event data, this implies that higher levels of (nonviolent) media attention are associated with lower likelihoods of reporting bias for repression events.[22]

[22] Although the substantive magnitudes of the coefficient estimates on the AP and AFP Reports variables are smaller in the case of the ICEWS models. This may imply that reporting biases are less severe in the ICEWS context, and(or) that the total AP/AFP media attention received by a country is simply less predictive of the specific reporting bias issues found within the ICEWS data, relative to the SCAD data.

Table A.18: Models of Reporting Bias in Africa 2012-2013

	SCAD	ICEWS	SCAD	ICEWS
	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)
	Multi-Source	Multi-Source	Multi-Source	Multi-Source
	Constant Pr	Constant Pr	W/ Cov	W/Cov
		Pr(Misclassification AP)		
$GDPpc_{t-1}$	.	.	-0.280	-0.215
			(0.168)	(0.067)
Pop_{t-1}	.	.	-0.268	-0.287
			(0.106)	(0.059)
$Demo_{t-1}$	.	.	-0.386	0.182
			(0.416)	(0.137)
AP Reports	.	.	-0.033	-0.009
			(0.008)	(0.002)
Constant	0.558	0.412	7.951	7.244
	(0.148)	(0.070)	(2.073)	(1.098)
		Pr(Misclassification AFP)		
$GDPpc_{t-1}$	.	.	-0.203	-0.452
			(0.172)	(0.070)
Pop_{t-1}	.	.	-0.057	-0.122
			(0.121)	(0.066)
$Demo_{t-1}$	.	.	0.350	0.108
			(0.402)	(0.136)
AFP Reports	.	.	-0.023	-0.011
			(0.006)	(0.003)
Constant	0.005	-0.650	3.332	5.470
	(0.181)	(0.103)	(2.359)	(1.066)

Note: $N = 1,092$. Values in parentheses are standard errors.

Validation

The above analyses suggest that machine and human coded event data will yield similar theoretical findings regarding the determinants of country-month African repression. Yet these analyses do not reveal whether the resultant predictions obtained from these models of African repression are comparable across our ICEWS and SCAD-based models. To evaluate this question, one needs reliable GSRs on African repression. In this case, we turn to the latent-country year measures of human rights protection estimated by Fariss (2014). As Bagozzi and Berliner (2017) note, "[w]ile there is no perfect variable to capture objective 'on-the-ground' human rights conditions, the most advanced option at present is Fariss's (2014) dynamic latent human rights protection measure" (14). Indeed, Fariss (2014) uses a variety of standards-based human rights sources[23] and event based repression data sources[24] within a dynamic item response theory (IRT) model to recover a latent measure of repression that (i) minimizes measurement error (e.g., reporting bias) issues associated with any specific dataset and (ii) accounts for changing standards of human rights accountability over time. Given the above points, we believe this latent measure to offer the best opportunity for validation of our models' predictions of repression.

To validate our ICEWS and SCAD based multi-source models in this manner, we specifically use the latent mean of countries' human rights scores for our sample from Fariss (2014, Version 2.4). Note that, on this measure, higher values imply better human rights performance (i.e., less repression). We then derive the in-sample predicted probabilities of repression for each country-month in our Africa sample from our (SCAD and ICEWS-based) repression stage estimates, separately for both the *multi-source with covariates* and *multi-source constant* specifications discussed above.[25] As both Fariss' latent human rights measure and our model-derived predicted probabilities are continuous, we examine the Pear-

[23] I.e., sources coded from annual Amnesty International and State Department human rights reports.

[24] Which are coded from a wide variety of historical, newspaper, newswire, and online sources, but which exclude the SCAD and ICEWS data used here.

[25] We use the repression stage estimates in this case as they provide us with our multi-source models estimated effects of each relevant covariate on repression, after accounting for reporting bias.

son product-moment correlations between our model predictions and Fariss's latent measure, and these correlations' associated t-values. We expect a negative correlation between our predicted probabilities and Fariss' measure, given that higher values on the former imply a higher likelihood of repression, whereas higher values on the latter imply less overall repression.

Our correlation results appear in Table A.19. Beginning with our *multi-source constant* specifications, we find that our SCAD- and ICEWS based models exhibit highly negative and statistically significant correlations with Fariss's measure. Moreover, each predicted probability exhibits a *very similar* correlation with Fariss's latent human rights protection scores: of -0.535 in the case of SCAD and -0.520 in the case of ICEWS. Thus, *multi-source constant* human- and machine-coded repression data yield predictions that are near-identical in their association with a set of plausible GSR data, although the anticipated correlation is slightly stronger in the case of our human-coded event data. Our finding for the *multi-source with covariates* case in Table A.19 underscore these conclusions. We find in this case that the inclusion of source specific misclassification predictors has improved the negative correlations between our predicted probabilities of repression and Fariss's latent human rights protection scores. Moreover, both our SCAD- and ICEWS-based predictions again yield near-identical correlations with these latent GSRs, of -0.593 in the case of SCAD and -0.590 in the case of ICEWS.

These similarities suggest that human and machine coded event data have comparable levels of external validity. As we discuss in the introduction to our Note, and in Bagozzi et al. (2016), this form of external validity is a necessary component to event data validation. Whereas much past research has focused on the internal validation of machine event data (i.e., the comparison of machine codings of specific texts to comparable human-codings of that same text), our analysis has thus provided one of the first external validations of machine coded event data—both relative to human coded event data and relative to an external (latent) gold standard source. In doing so, we find good reason to believe that machine

Table A.19: Correlation Coefficients with Latent Human Rights Protection Scores

	Pearson r	t-value
SCAD Pr(Repression) with Constant Pr	-0.535	-20.883
ICEWS Pr(Repression) with Constant Pr	-0.520	-20.121
SCAD Pr(Repression) with Covariates	-0.593	-24.299
ICEWS Pr(Repression) with Covariates	-0.590	-24.113

Note: $N = 1,092$

coded event data exhibits comparable external validity to human coded data.

Application 2: Colombian Human Rights Violations

Our second application examines instances of rebel and paramilitary violence against civilians in Colombia during the years 2000–2009. As mentioned briefly in the main paper, there are several compelling reasons for our focus on the Colombia case, and on violence against civilians. For one, the measurement of violence against civilians in Colombia has been an ongoing area of substantive and methodological focus for well over 30 years (Cingranelli and Pasquarello, 1985; Restrepo, Spagat and Vargas, 2006; CINEP, 2008; Ball et al., 2008; Lum et al., 2010). The prominence and scope of this body of research allows us to (1) benchmark our validation results, (2) substantively identify a number of sub-sample comparisons to isolate sources of reporting bias, and (3) ensure that our contributions have both scholarly and policy relevance. Unlike our analysis of repression in Africa above, this Colombia analysis additionally affords us the ability to examine reporting bias issues, and to validate machine and human event data, at fine-grained subnational levels. Subnational validation of this sort is likely to be of high interest to future researchers in this area, given the increasing shift towards grid-level (and/or municipality-level) analyses of conflict processes among quantitative conflict scholars.

Thanks in large part to previous research on violence against civilians, our focus on the Colombia case also ensures that—unique to this second application—we have access to a set of gold standard human rights violation (HRV) events through Colombia's Centro de Investigación y Educación Popular (CINEP; CINEP, 2008). The CINEP's HRV data are coded by event identification numbers attached to the official geopolitical division of Departments and Municipalities established by the *Departamento Administrativo Nacional de Estadística* (DANE). Each event reports the main participating actors in a hierarchical structure, with the victims of an event recorded by their number of deaths, injuries, disappearances, kidnappings, threats, attempts, arbitrary detention and forced recruitment. Each event is further classified according to more general categorical indicators for the type of HRV and type of civilian victimization. Importantly, the CINEP HRV data contain comprehensive information on rebel and paramilitary-perpetrated violence against civilians in Colombia for the years 1990-2009 on a monthly basis. These relational data — which record individual instances of human rights abuses at the municipality-month level — are unlikely to exhibit the reporting bias problems that are common to global (human- and machine-coded) event datasets. CINEP has been documenting the conflict in Colombia for over forty years, and has created an archive that is curated by librarians with an extensive collection of (Spanish language) national and regional Colombian newspapers and associated reports. This collection is the basis for CINEP's coding of HRV data, in addition to eye-witness and victim testimony, reports from NGOs, and government sources. As such, CINEP provides us with a GSR validation source that is generally unavailable for many other country-specific conflict applications.

A final justification for our choice of the Colombia case relates to the availability of overlapping human- and machine-coded event datasets for the purposes of comparison. The presence of a contemporary civil conflict in Colombia, and the recent growth of global event datasets more generally, together provide us with two well-documented human and machine-coded event datasets—the (machine coded) ICEWS data described above and the (human coded) Geo-located Event Dataset (GED; Sundberg and Melander, 2013)—for this specific conflict. Importantly for our specific validation goals, ICEWS and GED (i) each contain variables delineating the news source(s) that each dataset used to code a given event and (ii) exhibit considerable overlap in the specific news sources that each event dataset used to code Colombian HRV events. This news source overlap, as outlined for our African repression analysis above, allows us to use the aforementioned multi-source estimators to make subnational external validation comparisons of human and machine coded event data for the Colombian conflict. As noted in our Research Note's introduction[26] this form of external validation—along with internal validation—is a necessary component to the broader validation of machine coded event data.

Data Formatting

To perform these validation comparisons, we first must aggregate and combine the GED, ICEWS, and (CINEP) validation data for the case of Colombian HRVs. Note that our goals in this regard will ultimately be to compare the GED and ICEWS Colombian HRV data directly with the aid of the multi-source estimators described in our main Research Note and above. After doing so, we will use our held-out CINEP events to determine whether any discrepancies that arise in our misclassification-corrected GED and ICEWS HRV predictions can be attributed to better overall ground truth within the ICEWS (versus GED) data. However, before presenting these model-based comparisons in full, we must first discuss the formatting and aggregation choices that we use for our ICEWS, CINEP, and GED datasets. Following this, we present a series of descriptive statistics for our combined HRV data, before finally turning to our model-based comparisons and validation efforts.

Combining our ICEWS, GED, and CINEP datasets for the case of Colombian HRVs is not without its challenges. CINEP, ICEWS, and GED each exhibit different levels of spatio-temporal aggregation, have distinct definitions of what ultimately comprises an HRV event, and contain unique criteria for what constitutes HRV perpetrators and victims. These differences guarantee that any effort to combine all three datasets will have a some degree of error. What follows is a detailed discussion of our efforts to format and combine each of our

[26] And in Bagozzi et al. (2016).

event datasets in a manner that ensures that our retained HRV events are as comparable as possible across all three sources.

We start in this case by first describing our validation data. Our raw CINEP HRV data are aggregated to the municipality-year level, for the years 1990-2009. This sample-frame limits the end-year of our analysis to 2009. In addition, a majority of the municipality-year specific covariates that we include in our model-based validation efforts further below begin in the year 2000, which accordingly limits our sample's start-year to 2000. Our specific CINEP HRV data include directed rebel and paramilitary (source) to citizen (target) violence events. Directed dyad interactions of this sort (i) facilitate the comparison of the events with the directed dyadic event information contained in ICEWS and GED, and (ii) ensure that our analysis closely parallels the most common approach to event data coding and analysis within the field (i.e., dyadic relational interactions). The latter quality is a correspondence that many past event data validation experiments have ignored (King and Lowe, 2003). Within CINEP's HRV data, *source* actors are designated by the specific rebel or paramilitary group perpetrating an HRV event, and the *target* of each event can more simply be inferred to be a civilian or group of civilians. To combine these validation data with our anticipated event datasets, we first collapse CINEP's recorded rebel-perpetrated HRV events to the unique event-ID level. We then subset CINEP's events to only include actual instances of "material" human rights violations, rather than both material and verbal human rights violations.[27]

For the 2000-2009 period, we next aggregate all remaining CINEP HRV events to the municipality-year level.[28] We then dichotomize these municipality-year HRV event counts for use in our multi-source model-based validation efforts. As such, our final CINEP variable is equal to one for any municipality-year that experienced a rebel or paramilitary HRV, and zero otherwise. With the formatted CINEP HRV data in hand, we next formatted the

[27] That is, we remove all non-material violence events (e.g., threats), including categories such as 'Threatens', 'Recruitment', and 'Collective Threats,' which altogether constituted 75% of all rebel-perpetrated HRV events in CINEP for our years of analysis.

[28] Municipalities are Colombia's second administrative unit, with 1,102 municipalities in total.

World ICEWS data to correspond as closely as possible to our final CINEP events. As alluded to above, the actor and action designations within ICEWS do not precisely match those used by CINEP. In light of this, we again made every effort to be as comprehensive as possible in retaining all relevant source and target actors for comparison, while also not including any possibly questionable actor designations for our purposes. Within ICEWS, this entailed our treatment of actors designated as “rebel,” “separatist,” “insurgent,” and “unidentified sources” as source actors, and “general population,” “civilian,” and “social” as target actors. After identifying events occurring in Colombia between these source and target actors, we retained and aggregated (over location and year) only those events with the following CAMEO category 18 (ASSAULT) codes:

180: Use unconventional violence, not specified below
181: Abduct, hijack, or take hostage
182: Physically assault, not specified below
1821: Sexually assault
1822: Torture
1823: Kill by physical assault
183: Conduct suicide, car, or other non-military bombing, not specified below
1831: Carry out suicide bombing
1832: Carry out car bombing
1833: Carry out roadside bombing
184: Use as human shield
185: Attempt to assassinate
186: Assassinate

Before aggregating these ICEWS events, we applied a de-duplication criterion to ensure that only one event(-type) was recorded per day, source, and latitude-longitude coordinate.[29] After implementing this de-duplication routine, we retained only those events that were recorded by ICEWS from the following two newswire sources: Reuters and the Spanish international newswire agency Agencia EFE. These two newswire sources were chosen for comparison because they (i) encompassed a substantial number of ICEWS’ recorded HRV

[29] While ICEWS does very mild de-duplication at the coding stage—effectively eliminating duplicate stories bearing the same publisher, headline, and date—it still allows for some duplicate stories given (e.g.,) variation in headlines, which can lead to over-reporting of many events (Schrodt, 2015, 12).

events and (ii) similarly encompass a substantial number of the HRV events within the GED data discussed below. For *each* newswire specific sample of HRV events, we then aggregated that newswire source's ICEWS HRV events to the municipality-year level, and created a dichotomized indicator of these ICEWS-derived HRV event-measures for both Reuters-coded events and EFE-coded events. We next matched our ICEWS' (and our subsequent event datasets') events to maps of Colombia's municipalities via latitude-longitude coordinates. To do so, we first standardized the spatial reference of all datasets by projecting event point data of each event onto the World Geodetic Survey (WGS84) geographic coordinate system[30] via latitude-longitude coordinates. All event data points were then translated to, and merged to our municipalities and departments based upon, a two-dimensional Mercator Auxiliary Sphere.[31]

Importantly, the above ICEWS data aggregation decisions retained only a small subset of all ICEWS HRV events for Colombia. At the Colombian municipality-year level for the years 2000-2009, a total of 63 distinct news stories were identified as containing relevant HRV events within our ICEWS data. The EFE and Reuters-specific events that we retained from this full sample constituted only 1,738 (41%) of all 4,275 HRV events recorded by ICEWS from all relevant news sources. As was the case for our African repression analysis, our retained Colombian ICEWS HRV events exhibit underreporting *by design.* While this maximizes our abilities to (i) leverage the misclassification-estimators discussed earlier and (ii) compare our ICEWS findings to the GED data and findings discussed below; it ensures that our resultant events will poorly approximately our gold standard CINEP cases (which are coded from a multitude of sources) relative to what could be obtained from the full ICEWS data. Note that—based on our Monte Carlo extension 1—this also again implies that the multi-source models considered below will potentially yield biased and imprecise

[30] Which uses three-dimensional spherical surface to define each point location on the earth.

[31] We specifically used the WGS 1984 Web Mercator Auxiliary Sphere projected coordinate system. This was the best projection available for our application because it uses a spheroid, rather than perfect sphere, for its earth model, thereby making it more efficient in aligning local data. By matching the projection and geographic coordinate system we are able to say with confidence that the locations are properly located within our study area.

estimates for the models below; yet these estimates can still be anticipated to be superior to those obtained from single source (e.g., probit) models.

We next formatted our GED dataset in a comparable manner to the ICEWS formatting steps described above. Recall that GED is a (near-global) human-coded event dataset that draws on both news(wire) sources and NGO reports for its coding of individual events. For the Colombia case, we retained all nonstate (rebel or paramilitary) perpetrated HRVs against civilians (i.e., "one-sided violence") within the GED data for the years mentioned above, while taking care to exclude any HRVs perpetrated explicitly by Colombian drug cartels. We then split these events into two datasets that separately recorded (i) the GED based HRVs that were coded from stories appearing in Reuters, and (ii) the GED HRVs that were coded from stories appearing in EFE. As above, this retains a small subset of the total Colombian HRVs included in GED: only 48% of all GED Colombian HRVs were coded from these two sources for the 2000-2009 period. With this caveat in mind, our two GED datasets were then merged to Colombian municipality-year maps for the 2000-2009 period, while taking care to omit any GED events whose levels of geocoding accuracy were determined to be too ambiguous to fit within this particular administrative level. Finally, we dichotomized these GED HRV events, for both newswire-specific samples, at the municipality-year level. After our formatting and aggregation tasks were complete, we combined all HRV measures discussed above into a single municipality-year dataset covering the years 2000-2009.

Dichotomized HRV Measure Comparisons

This subsection compares our dichotomized HRV indicators using summary statistics, bivariate comparisons, and binary classification criteria. Turning first to Tables A.20-A.23, we find that our event datasets (i.e., ICEWS and GED) do a fairly poor job in accurately identifying the municipality-years that experience at least one HRV violation according to CINEP. Recall, however, that the retained ICEWS and GED HRV events only correspond to a small subset of all HRV events recorded in each of these datasets. In order to conduct a meaningful comparison of the validity of machine and human coded event data, we retained only those events within each dataset that were recorded from Reuters and EFE.

With the above points in mind, the confusion matrices presented in Table A.21 indicate that comparisons of our dichotomized (news source-specific) event datasets to CINEP at the municipality-year level yield more false positives than true positives within ICEWS (Reuters and EFE) and the EFE-specific GED data. The Reuters-specific GED data exhibits slightly more true positives than false positives. For both ICEWS and GED, the EFE news source identifies more CINEP events than does Reuters (albeit with more false positives as well). Likewise, GED exhibits a slight edge over ICEWS in terms of total true positives, whereas ICEWS identifies more total events for both Reuters ($117 > 81$) and EFE ($228 > 161$) than does GED. Hence, taken together, the confusion matrices presented in Table A.21 suggest that the Reuters and EFE events that are recorded within ICEWS and GED are each fairly similar to one another in their (fairly poor) aggregate approximations of our gold standard CINEP HRVs.

The confusion matrices in Table A.22 allow us to more directly compare our news source-specific event datasets to one another. The confusion matrices in this table specifically compare the (i) Reuters-specific ICEWS data to the EFE-specific ICEWS data (Table A.22a), (ii) the Reuters-specific GED data to the EFE-specific GED data (Table A.22b), (iii) the Reuters-specific HRVs across ICEWS and GED (Table A.22c), and (iv) the EFE-specific HRVs across ICEWS and GED (Table A.22d). Of most relevance to our anticipated multi-source models, Table A.22a indicates that our two news source-specific ICEWS datasets each uniquely code 191 (EFE) and 80 (Reuters) municipality-years as experiencing at least one HRV; and jointly code 37 municipality-years as experiencing at least one HRV. For the GED's HRVs (Table A.22b) we find 143 (EFE) and 63 (Reuters) uniquely-identified municipality-years, and 18 municipality-years jointly coded as experiencing at least one HRV by both Reuters and EFE. Among both GED and ICEWS, we thus have high levels of reporting bias and low levels of two-source overlap, since each source appears to capture a relatively low number of that dataset's other source's municipality-year HRVs, and ostensibly, of HRVs overall. Per our Monte Carlo extension one, this leads us to anticipate relatively imprecise estimates for the various probit and multi-source models assessed below.

Tables A.22c-A.22d suggest that for each news source (i.e., Reuters or EFE) the HRV events recorded in ICEWS and GED are largely distinct in their identified municipality-years (with only 11-24 cases of overlap). Hence, while the (Reuters and EFE-specific) ICEWS and GED data each appear to contain a similar number of events, our source-specific event datasets exhibit substantial disagreement in *which* municipality-years experience HRVs, and which do not, for Colombia during the years 2000-2009.

Our final set of pairwise comparisons again treats the dichotomized CINEP HRV data as "truth," and calculates a series of classification statistics (defined on page 50 below) to evaluate how well each source-specific event dataset recovers our CINEP HRV municipality-year cases. These classification statistics are reported in Table A.23 and slightly favor GED over ICEWS (and EFE over Reuters) with respect to the accurate classification of the (dichotomized) CINEP data. Even so, we can see in Table A.23 that all news source specific event datasets perform poorly in classifying actual CINEP HRVs (via sensitivity), but generally perform well on specificity and overall accuracy due to the preponderance of zero-cases (i.e., non events) across all datasets. If we consider CINEP as our GSRs, these results accordingly suggest that the Reuters and EFE-specific events contained in ICEWS and GED[32] are individually fairly poor approximations of HRV "truth" at the municipality-year level. Each comparison dataset also appears to exhibit somewhat distinct deficiencies in these regards. However, on the whole, the Reuters and EFE-specific HRV events contained in ICEWS and GED appear to be fairly similar in their (in)abilities to accurately classify our CINEP HRV cases at the municipality-year level of aggregation.

[32] Which, again, are only a small fraction of all Colombia HRVs recorded in these two datasets, given the many additional news sources that each dataset codes, but which we omit here.

Table A.20: Dichotomized Municipality-Year Summary Statistics (2000-2009)

	Mean	Stdev	Min	Max	Corr w/ CINEP	Total Conflict Cases
CINEP	0.092	0.289	0	1	1.000	1,029
ICEWS (Reuters)	0.010	0.102	0	1	0.062	117
ICEWS (EFE)	0.020	0.141	0	1	0.097	228
GED (Reuters)	0.007	0.085	0	1	0.130	81
GED (EFE)	0.014	0.119	0	1	0.149	161

Note: $N = 11,220$ (11,22 municipalities × 10 years)

Table A.21: Dichotomized Municipality-Year Confusion Matrices (CINEP Comparisons)

(a) CINEP Vs. ICEWS (Reuters)

		ICEWS		
		0	1	Total
CINEP	0	10,105	86	10,191
	1	998	31	1,029
	Total	11,103	117	11,220

Pearson $\chi^2 = 42.601$, $P < 0.001$

(b) CINEP Vs. ICEWS (EFE)

		ICEWS		
		0	1	Total
CINEP	0	10,028	163	10,191
	1	964	65	1,029
	Total	10,992	228	11,220

Pearson $\chi^2 = 104.475$, $P < 0.001$

(c) CINEP Vs. GED (Reuters)

		GED		
		0	1	Total
CINEP	0	10,153	38	10,191
	1	986	43	1,029
	Total	11,139	81	11,220

Pearson $\chi^2 = 188.893$, $P < 0.001$

(d) CINEP Vs. GED (EFE)

		GED		
		0	1	Total
CINEP	0	10,102	89	10,191
	1	957	72	1,029
	Total	11,059	161	11,220

Pearson $\chi^2 = 247.811$, $P < 0.001$

Table A.22: Dichotomized Municipality-Year Confusion Matrices (Event Data Comparisons)

(a) ICEWS (EFE) Vs. ICEWS (Reuters)

		EFE		
		0	1	Total
Reuters	0	10,912	191	11,103
	1	80	37	117
	Total	10,992	228	11,220

Pearson $\chi^2 = 520.064$, $P < 0.001$

(b) GED (EFE) Vs. GED (Reuters)

		EFE		
		0	1	Total
Reuters	0	10,996	143	11,139
	1	63	18	81
	Total	11,059	161	11,220

Pearson $\chi^2 = 249.271$, $P < 0.001$

(c) ICEWS (Reuters) Vs. GED (Reuters)

		GED		
		0	1	Total
ICEWS	0	11,033	70	11,103
	1	106	11	117
	Total	11,139	81	11,220

Pearson $\chi^2 = 124.283$, $P < 0.001$

(d) ICEWS (EFE) Vs. GED (EFE)

		GED		
		0	1	Total
ICEWS	0	10,855	137	10,992
	1	204	24	228
	Total	11,059	161	11,220

Pearson $\chi^2 = 136.005$, $P < 0.001$

Table A.23: Dichotomized Municipality-Year Classification Statistics (2000-2009)

	False Positive Rate	False Negative Rate	Sensitivity	Specificity	F1 Score	Total Accuracy
ICEWS (Reuters)	0.01	0.97	0.03	0.99	0.05	0.90
ICEWS (EFE)	0.02	0.94	0.06	0.98	0.10	0.90
GED (Reuters)	0.004	0.96	0.04	0.996	0.08	0.91
GED (EFE)	0.01	0.93	0.07	0.99	0.12	0.91

Note: $N = 11,220$ (11,22 municipalities × 10 years)

Multi-source Model Comparisons

Our model-based comparisons of the GED and ICEWS Colombian HRV data proceed in a similar fashion to the African repression analysis summarized above. Here we estimate three specific models for each HRV dependent variable (i.e., for our GED and ICEWS-derived HRV dependent variables): (i) a standard probit model, (ii) a *multi-source constant* model and (iii) a *multi-source with covariates* model. For the HRV stage of each model, we include three plausible (municipality-year level) variables known to affect rebel and paramilitary violence within the context of Colombia (Angrist and Kugler, 2008; Richani, 2013; Holmes and Gutiérrez de Piñeres, 2014; Ibánez, 2009; Holmes et al., 2017): (logged) population,[33] the percentage of a municipality with forest cover (as a proxy for both remoteness and for the types of jungle-cover that are conducive to rebel and paramilitary operations in Colombia),[34] and the percentage change in population in a given municipality from the previous year to the present year (to approximate migration pressures).[35] The two latter percentage-based variables are measured on proportion (i.e., 0-1) scales. We also add each of these measures to the Reuters and EFE-specific misclassification stages of our *multi-source with covariates* models, along with an additional measure of remoteness: the logged distance from each municipality's centroid to the capital Bogotá.[36]

As alluded to, our rationale for including the latter distance variable, along with percentage forest cover, in the misclassification stage of our model follows from the notion that more remove locations may receive less press attention, and lower recognition of HRVs. Indeed,

[33] To construct municipality-year level estimates of Colombia's population 2000-2009, municipality level population data were obtained for the years 1985, 1993, and 2005 from Colombia's Departamento Administrativo Nacional de Estadística (DANE):`https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/series-de-poblacion`. These data were then interpolated to the yearly level using natural splines with pivot information fitted on June 30th, 1985, 1993 and 2005.

[34] Derived from the Consortium for Spatial Information's digital elevation model.

[35] Population change is based on percentage changed that were derived the same DANE estimations mentioned above, for the years 2000-2009.

[36] To construct this measure, we converted each municipality polygon's centroid x- (longitude) and y-coordinates (latitude) into a new shapefile of centroid points. Using a generally accepted latitude-longitude point for Bogotá, we next measured the distance in meters from each centroid point to Bogotá to determine the final distance of municipality and department to the capital. We then transformed these measures by taking its natural log to address issues of skewness, and in order to ensure reasonably scaled coefficient estimates.

there is ample existing evidence to suggest that issues of reporting bias and non-detection in media-derived become far more severe as one moves into the more sparsely populated and remote areas of a conflict-afflicted country (Davenport and Ball, 2002; Weidmann, 2015). We anticipate that increases in a municipality's distance from Bogotá (and forest cover) will lead to higher reporting biases. We also draw upon this same rational in justifying our inclusion of logged population and population change in the misclassification stages of our *multi-source with covariates* model—as well as upon the Africa analysis in Cook et al. (2017), and our extensions of this analysis above, which reveal that population is generally a negative and statistically significant predictor of misclassification in the African repression case.

The results from all models mentioned above appear in Tables A.24-A.25. Beginning with the HRV-outcome results in Table A.24, we find stable results for virtually all coefficient estimates across our SCAD and ICEWS specifications. Forest cover is consistently positive across all specifications, suggesting increased forest cover to be associated with a higher likelihood of a HRV at the municipality-year level, though it is not statistically significant in the standard probit specifications. Intuitively, logged population is consistently positive and statistically significant across all HRV models and event data analyzed in Table A.24, though its estimated effects attenuate in size as one moves to the more specified *multi-source with covariates* models. Finally, population change exhibits perhaps the most notable change in Table A.24, wherein we find population change to be negative but not statistically significant in the probit and *multi-source constant* models, but positive in the *multi-source with covariates* models.[37] Together the results in Table A.24 suggest—that no matter whether one chooses to use a probit, *multi-source constant* model, or *multi-source with covariates* model—the findings obtained (in terms of significance, sign, and magnitude) will be similar when using either human- or machine-coded HRV event data.

[37] And statistically significant at the $p < .10$ level in the case of the ICEWS *multi-source with covariates* model.

Table A.24: Models of HRVs in Colombia 2000-2009

	GED	ICEWS	GED	ICEWS	GED	ICEWS
	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)
	Probit	Probit	Multi-Source	Multi-Source	Multi-Source	Multi-Source
			Constant Pr	Constant Pr	W/ Cov	W/Cov
Forest	0.621	0.361	0.783	0.412	3.437	1.018
	(0.129)	(0.124)	(0.174)	(0.144)	(0.657)	(0.443)
Log Pop.	0.312	0.391	0.401	0.447	0.235	0.270
	(0.024)	(0.022)	(0.039)	(0.029)	(0.061)	(0.043)
Pop. Change	-1.009	-1.217	-1.005	-1.211	4.409	9.865
	(1.576)	(1.418)	(2.131)	(1.809)	(4.753)	(5.974)
Constant	-5.265	-5.885	-5.693	-6.131	-4.209	-4.125
	(0.252)	(0.234)	(0.372)	(0.292)	(0.700)	(0.521)

Note: $N = 9,451$. Values in parentheses are standard errors.

We also find fairly consistent results across our ICEWS and GED models within the misclassification stages of our *multi-source with covariates* models (see Table A.25). For example, across both the Reuters and EFE-specific misclassification equations in Table A.25, we consistently find Forest Cover to be a positive predictor of misclassification within our ICEWS and GED models; although the coefficient estimate on Forest Cover is not statistically significant in the case of the ICEWS-Reuters equation. This null finding notwithstanding, these results intuitively imply that municipalities with higher forest cover tend to experience higher levels of misclassification and thus reporting biases. Next, and similar to our findings for Africa repression above, we find that Log Population is a statistically significant negative predictor of misclassification for both news sources, and event datasets. This again is intuitive, in its implying that more populated (and likely urban) areas are less likely to experience reporting biases. Increases in population, on the other hand, are generally associated with increases in misclassification rates,[38] perhaps due to the broader social disruptions caused by inward migration (after controlling for total population levels within each municipality), or due to this variable's proxying for agricultural areas. Finally, we generally find inconsistent results for logged distance (in terms of significance, sign, and magnitude) across our media sources, and event datasets, suggesting perhaps that distance is a poor proxy for remoteness in this context, especially after one has accounted for population-based factors and forest cover.

[38] Though this effect is not always statistically significant.

Table A.25: Models of Reporting Bias in Colombia 2000-2009

	GED	ICEWS	GED	ICEWS	GED	ICEWS
	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)	(Human-Coded)	(Machine-Coded)
	Probit	Probit	Multi-Source	Multi-Source	Multi-Source	Multi-Source
			Constant Pr	Constant Pr	W/ Cov	W/Cov
Pr(Misclassification Reuters)						
Forest	.	.	.	.	1.449	0.055
					(0.368)	(0.364)
Log Pop.	.	.	.	.	-0.252	-0.281
					(0.062)	(0.046)
Pop. Change	.	.	.	.	12.074	15.892
					(6.665)	(6.187)
Log Distance	.	.	.	.	-0.140	0.032
					(0.099)	(0.076)
Constant	.	.	1.150	0.837	5.415	3.693
			(0.122)	(0.090)	(1.419)	(1.184)
Pr(Misclassification EFE)						
Forest	.	.	.	.	1.581	0.805
					(0.364)	(0.404)
Log Pop.	.	.	.	.	-0.198	-0.340
					(0.065)	(0.048)
Pop. Change	.	.	.	.	1.693	14.585
					(4.927)	(6.379)
Log Distance	.	.	.	.	-0.105	-0.266
					(0.083)	(0.073)
Constant	.	.	0.674	0.302	4.026	7.455
			(0.145)	(0.108)	(1.288)	(1.118)

Note: $N = 9,451$. Values in parentheses are standard errors.

Validation

We validate our multi-source models and results against our gold standard HRV data, which was discussed both above and in our main Research Note. Recall that the GSRs employed here were drawn from Colombia's CINEP data (CINEP, 2008). Like the Colombia event data analyzed above, CINEP's HRV data contain comprehensive information on rebel and paramilitary-perpetrated violence against civilians in Colombia for our event types of interest. As mentioned previously, these CINEP data are unlikely to exhibit the reporting bias problems that are common to global (human- and machine-coded) event datasets. This in large part because CINEP has been documenting the conflict in Colombia for over forty years, and has created an archive that is curated by librarians with an extensive collection of (Spanish language) national and regional Colombian newspapers and associated reports. This collection is the basis for CINEP's coding of HRV data, in addition to victim testimony, NGO reports, and government sources. As we mention in our main Research Note, these features provide us with a GSR validation source (i.e., CINEP) that is generally unavailable for country-specific conflict applications, and the bivariate comparisons presented above help to underscore this point.

To perform our validation exercises, we first extract the (misclassification-adjusted) HRV predictions from our Colombia-specific two-source models. We do so by generating the predicted probabilities of HRV for each municipality-year in our sample, separately from each of our multi-source estimators' HRV stage estimates. We then compare these predictions to our binary CINEP records of municipality-year HRVs using areas under the receiver operating characteristic curve (AUCs) and areas under the precision-recall curve (AUC-PRs). Given the relative rarity of the events of interest, we favor the latter metric over the former (Ward and Beger, 2017). The results of these comparisons are reported in Table A.26.

Turning to Table A.26, we can first note that all model predictions do a poor-to-modest job of classifying our binary CINEP HRV data, with AUCs ranging from 0.649 to 0.673. This relatively poor overall performance in classifying our CINEP events is also reflected in our AUC-PRs, which range from 0.166 to 0.184 in Table A.26. However, recall that the ICEWS and GED HRV measures used here represent only a tiny fraction of all HRV events recorded in ICEWS and GED, given our subsetting of these data to only include events coded by Reuters and EFE. As such, our models' poor overall classification of the CINEP cases is to be expected, and is in no way indicative of the overall quality of the ICEWS or GED data. What is instead most relevant to the comparisons at hand are the comparisons of AUC-PRs (or AUCs) between our GED and ICEWS based models, for a given model specification. Here we find that in each and every case, our ICEWS and GED-based models perform comparably in classifying our held-out CINEP events. For example, in the *multi-source constant* specifications reported in Table A.26 which are the best performing models classification-wise, we find that our GED models yield an AUC of 0.673 whereas our ICEWS model yields an AUC of 0.661. These are effectively identical AUCs when rounded to the second decimal point. The AUC-PRs for these constant-specifications are highly similar as well, and range from 0.174 (ICEWS) to 0.184 (GED), thus slightly favoring the GED data on this metric.

Turning to our *multi-source with covariates* model specifications, we again find very similar AUCs and AUC-PRs across our GED and ICEWS models. In the case of AUCs, we find here that our GED model provides an AUC of 0.665, whereas our ICEWS model yields a slightly worse but still very comparable AUC (of 0.649). However, these patterns are reversed in the case of the AUC-PRs reported in the bottom half of Table A.26. In this case, we find that our ICEWS and GED AUC-PRs are again very similar, but now slightly favor the ICEWS data (AUC-PR= 0.178) over the GED data (AUC-PR= 0.166). In sum, we find in Table A.26 that each pairing of AUC-PRs (and each pairing of AUCs) is highly similar across our GED and ICEWS-based data on models. This suggests that Colombian municipality-year models employing either the GED or ICEWS data are comparable in their abilities to predict CINEP HRVs—and in most cases these predictions are effectively identical. Hence, for analyses of subnational HRVs in Colombia, our application we have obtained similar

Table A.26: Classification of CINEP HRVs

	AUC	AUC-PR
GED Pr(HRV) with Constant Pr	0.673	0.184
ICEWS Pr(HRV) with Constant Pr	0.661	0.174
GED Pr(HRV) with Covariates	0.665	0.166
ICEWS Pr(HRV) with Covariates	0.649	0.178

Note: $N = 9,451$

substantive conclusions *and* similar predictive accuracy when utilizing machine or human coded event data. These findings—along with those of our Africa application—help to underscore the external validity of machine coded event data relative to human coded data. External validation of this form represents a critical, and often overlooked, component to the validation of machine coded event data (Bagozzi et al., 2016). In these regards, we believe that the applications discussed above represent novel contributions to the measurement and validation of (machine-coded) political event data.

Classification Formulas

$$Sensitivity = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Negatives} \tag{A.3}$$

$$Specificity = \frac{number\ of\ True\ Negatives}{number\ of\ True\ Negatives + number\ of\ False\ Positives} \tag{A.4}$$

$$False\ Positive\ Rate = 1 - \frac{number\ of\ True\ Negatives}{number\ of\ True\ Negatives + number\ of\ False\ Positives} \tag{A.5}$$

$$False\ Negative\ Rate = \frac{number\ of\ False\ Negatives}{number\ of\ Fale\ Negatives + number\ of\ True\ Positives} \tag{A.6}$$

$$Pos.\ Predictive\ Value\ (PPV) = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Positives} \tag{A.7}$$

$$F1\ Score = 2 * \frac{PPV * Sensitivity}{PPV + Sensitivity} \tag{A.8}$$

$$Total\ Accuracy = \frac{number\ of\ True\ Positives + number\ of\ True\ Negatives}{number\ of\ cases} \tag{A.9}$$

References

Angrist, Joshua D. and Adriana D. Kugler. 2008. "Rural Windfall or a New Resource Curse? Coca, Income, and Civil Conflict in Colombia." *Review of Economics and Statistics* 90(2):191–215.

Bagozzi, Benjamin E. and Daniel Berliner. 2017. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of U.S. State Department Human Rights Reports." *Political Science Research and Methods* .

Bagozzi, Benjamin E., Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim and Agustin Palao. 2016. "External Validation of Event Data." Paper presented at the Annual Meetings of the APSA.

Ball, Patrick, Tamy Guberek, Daniel Guzmán, Amelia Hoover and Meghan Lynch. 2008. "Assessing Claims of Declining Lethal Violence in Colombia." Working Paper of the Human Rights Program of the Benetech Initiative.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2016. "ICEWS Coded Event Data." `http://dx.doi.org/10.7910/DVN/28075`, Harvard Dataverse.

CINEP. 2008. "Marco Conceptual: Banco de Datos de Derechos Humanos y Violencia Política." Centro de Investigación y Educación Popular.

Cingranelli, David L. and Thomas E. Pasquarello. 1985. "Human Rights Practices and the Distribution of U.S. Foreign Aid to Latin American Countries." *American Journal of Political Science* 29(3):539–563.

Cook, Scott J., Betsabe Blas, Raymond J. Carroll and Samiran Sinha. 2017. "Two Wrongs Don't Make a Right: Addressing Underreporting in Binary Data from Multiple Sources." *Political Analysis* 25(2):223–240.

Davenport, Christian and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–316.

Hausman, Jerry A., Jason Abrevaya and Fiona M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-Response Setting." *Journal of Econometrics* 87(2):239–269.

Hendrix, Cullen S. and I. Salehyan. 2012. "Social Conflict in Africa Database. Version 3.0. Codebook and coding procedures.".

Hendrix, Cullen S. and I. Salehyan. 2015. "No News is Good News? Mark and Recapture for Event Data When Reporting Probabilities are Less than One." *International Interactions* 42(2):392–406.

Hendrix, Cullen S. and Idean Salehyan. 2016. "A House Divided Threat Perception, Military Factionalism, and Repression in Africa." *Journal of Conflict Resolution* .

Hendrix, Cullen S., Idean Salehyan, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull and Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38(4):503–511.

Holmes, Jennifer, Agustin Mendizabal, David De La Fuente, Kristjan Mets, Alvaro Cárdenas, Liliana Dávalos and Dolors Armenteras. 2017. "Identifying municipal risk factors for leftist guerrilla violence in Colombia." Working Paper.

Holmes, Jennifer S. and Sheila Amin Gutiérrez de Piñeres. 2014. "Violence and the State: Lessons from Colombia." *Small Wars & Insurgencies* 25(2):372–403.

Ibánez, Ana María. 2009. "Forced displacement in Colombia: magnitude and causes." *The Economics of Peace and Security Journal* 4(1):48–54.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for Interational Conflict Data with Preformance as Good As Human Coders." *International Orgaization* 57(3):617–642.

Lum, Kristian, Megan Price, Tamy Guberek and Patrick Ball. 2010. "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007." *Statistics, Politics, and Policy* 1(1).

Restrepo, Jorge A., Michael Spagat and Juan F. Vargas. 2006. "The Severity of the Colombian Conflict: Cross-Country Datasets Versus New Micro-Data." *Journal of Peace Research* 43(1):99–115.

Richani, Nazih. 2013. *Systems of Violence: The Political Economy of War and Peace in Colombia, Second Edition*. Albany, NY: SUNY Press.

Schrodt, Philip A. 2015. "Comparison Metrics for Large Scale Political Event Data Sets." Paper presented at the Text as Data meetings, New York University, 16-17 October 2015.

Schrodt, Philip A., Deborah J. Gerner and Omur Yilmaz. 2009. *International Conflict Mediation: New Approaches and Findings*. New York: Routledge chapter Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World.

Sundberg, R. and E. Melander. 2013. "Introducing the UCDP georeferenced event dataset." *Journal of Peace Research* 50(4):523–532.

Ward, Michael D. and Andreas Beger. 2017. "Lessons from Near Real-time Forecasting of Irregular Leadership Changes." *Journal of Peace Research* 54(2):141–156.

Weidmann, Nils B. 2015. "On the Accuracy of Media-based Conflict Event Data." *Journal of Conflict Resolution* 59(6):1129–1149.