Two Wrongs Make a Right:

Addressing Underreporting in Binary Data from Multiple Sources*

Scott J. Cook[1], Betsabe Blas[2], Raymond J. Carroll[3], and Samiran Sinha[4]

[1] *Department of Political Science, Texas A&M University*

[2] *Department of Statistics, Federal University of Pernambuco*

[3] *Department of Statistics, Texas A&M University and*
School of Mathematical Sciences, University of Technology Sydney

[4] *Department of Statistics, Texas A&M University*

*For their helpful comments and suggestions, thanks to Florian Hollenbach, Idean Salyehan, Nils Weidmann, the reviewers, and the editor(s). Replication materials are available online at Cook et al. (2016). All inquiries should be sent to the corresponding author at sjcook@tamu.edu

Abstract

Media-based event data – i.e., data comprised from reporting by media outlets – are widely used in research in political science. However, events of interest (e.g., strikes, protests, conflict, etc.) are often underreported by these primary and secondary sources, producing incomplete data that risks inconsistency and bias in subsequent analysis. While general strategies exist to help ameliorate this bias, these methods do not make full use of the information often available to researchers. Specifically, much of the event data used in the social sciences is drawn from multiple, overlapping news sources (e.g., Agence France-Presse, Reuters, etc.). Therefore, we propose a novel maximum likelihood estimator that corrects for misclassification in data arising from multiple sources. In the most general formulation of our estimator, researchers can specify separate sets of predictors for the true-event model and each of the misclassification models characterizing whether a source fails to report on an event. As such, researchers are able to accurately test theories on both the causes of and reporting on an event of interest. Simulations evidence that our technique regularly outperforms current strategies that either neglect misclassification, the unique features of the data-generating process, or both. We also illustrate the utility of this method with a model of repression using the Social Conflict in Africa Database.

1 Introduction

Media-based event data – i.e., data comprised from newspaper, television, or web-based accounts – are widely used in research in political science, economics, sociology, and geography. Earl et al. (2004) details the centrality of these data in the research on collective action – e.g., racial violence, agrarian protest, social movements – arguing that for such issues there is simply "no other alternative available." In comparative politics, these data are used in the study of coups, demonstrations, natural disasters, elections, terrorism, and other forms of political violence. Research on intrastate conflict, in particular, frequently uses media-reported data to assess subnational variation in violence and gain greater leverage on the mechanisms which produce conflict (Weidmann, 2016). International Relations research has long used event data to detail the wide array of political interactions between countries, from disputes to diplomacy (Schrodt and Gerner, 1994). Media-based measures can be particularly important when political actors wish to conceal their behavior – as frequently occurs, in areas as diverse as human rights violations and Chinese development finance to Africa (Strange et al., 2013). Studies relying on media-based data will likely to continue to grow, as these data become more abundant and research questions become increasingly granular (Schrodt, 2012).[1]

Despite the broad and frequent use of these data, however, Woolley (2000, p. 177) notes that "[m]edia sources may not provide good data even though the data are sometimes easily obtained." In particular, concerns over measurement error are often raised, as news sources do not report all events.[2] The extent of this underreporting is difficult to assess – as we rarely possess true population rates – yet, some studies have estimated that nearly half of true events go unreported. This incomplete or selective reporting is often attributed to either a lack of opportunity during news gathering (e.g., the distance of a reporter to an event) or willingness during news reporting (e.g., the severity of an event, perceived audience demand, political bias of the media outlet), resulting in the systematic exclusion of particular events

[1] Schrodt (2012) provides a partial list of event data projects currently used in political science.

[2] These issues are widely recognized in affected literatures. For example, the literature on civil conflict has had numerous recent discussions on potential underreporting (Weidmann, 2014).

from media and resultant data.[3] Issues that are even further compounded when reporting on any illegal, illicit, or otherwise socially disapproved behavior – from corruption (e.g., vote buying, bribes, etc.) to conflict (e.g., coups, repression, etc.) – that are often of interest in social science. Consequently, data from these reports are misclassified – wherein some true events are coded as zeros – and analysis suffers from bias, with inferences which are sensitive to the choice of the source. Despite wide knowledge of these limitations, the threat of reporting bias is often ignored in applied research utilizing event count data. Where solutions have been proposed, researchers are advised to "triangulate" their data – draw from multiple news sources – to reduce the bias introduced from any one source, or to correct for the bias in estimation by modeling the misclassification (Hug and Wisler, 1998; Hug, 2003, 2009).

While either of these approaches is preferable to ignoring possible misclassification outright, we argue that neither is able to fully exploit all of the information commonly possessed by researchers with media-based event data. Namely, no strategy allows researchers to use multiple sources of reporting *and* misclassification-robust estimation. Briefly summarizing existing approaches, where only one source of (suspected to be incomplete) information is possible – that is, triangulation cannot be accomplished – we agree that researchers should use misclassification-robust methods (Copas, 1988; Carroll and Pederson, 1993; Hausman, Abrevaya and Scott-Morton, 1998). Where instead multiple sources of information are available, we agree that this should also be leveraged. However, the aggregation of these sources only attenuates, not eliminates, underreporting in the data, meaning additional statistical remedies for misclassification should still be employed. Yet, as none of the current misclassification-robust estimation strategies are derived for this kind of multi-source data-generating process, their application to these data will result in bias or loss of efficiency. Therefore, we propose a misclassification-robust maximum likelihood estimator for multiple sources of data, allowing researchers to estimate the extent of misclassification in each source, and obtain the correct estimates of event of interest. We further generalize our estimator to allow researchers to estimate models where misclassification is dependent upon covariates.

While the focus of our discussion centers largely on media-sourced data, our method is

[3] See Earl et al. (2004) for a more comprehensive discussion on media reporting.

more general than this suggests. Any time a researcher possesses multiple sources of information (ex. U.S. Department of State vs. United Nations reports) on a qualitative outcome of interest our estimator may be of use.[4] In the next section we briefly summarize the implications of response measurement error in discrete-outcome models. Next, we introduce our multi-source misclassification estimator. Following that, we outline and present results from Monte Carlo simulations evaluating our estimator against plausible alternatives. Anticipating our findings, we show that neglecting misclassification, by estimating a standard probit, *never* recovers the true effect estimate when there is differential misclassification, indicating the clear need to adopt corrections such as ours.[5] We then apply our method to a model of repression using the Social Conflict in Africa Database. Finally, we detail and discuss extensions of our approach to more general applied settings – e.g., scaling the estimator for data with many sources, analyzing non-binary qualitative outcomes, underreporting in sample selection – before concluding.

2 Measurement Error as Misclassification

Concern over measurement tends to focus predominately on error in predictors rather than error in responses. In part, this is because classical measurement error in the outcome of a linear regression 'only' increases the variability of fitted lines without otherwise causing bias to the slope estimates. As discussed in Carroll et al. (2006), this is not the case in discrete-outcome models, where measurement error *is* misclassification; risking not only a loss of precision but also bias in effects estimates. This can take two forms:

[4] As a minimal example, Trumbore and Woo (2014) analyze the conditions which lead states to become narcotic producers or transit platforms. In their analysis, they utilize data culled from the annual International Narcotics Control Strategy Reports published annually by the Bureau for International Narcotics and Law Enforcement Affairs of the United States Department of State. Using our method they could have supplemented this analysis to include the annual World Drug Report of United Nations Office of Drug Control as a second source and then estimated misclassification probabilities – with different reporting rates a function of the distinct political goals/aims of the two actors (i.e., US vs. UN) generating the reports. As this example shows, our model is appropriate whenever researchers have multiple sources of data from which an indicator of some event can be derived.

[5] The coverage probabilities for probit are 0.0% in our simulation experiment where there is differential misclassification of approximately 35% in one source and 20% in the other. The details of this analysis are provided below.

1. Non-differential misclassification – when the observed outcome is independent of the covariates conditional upon the actual outcome, that is, the event predictors do not also predict classification – induces severe attenuation bias in parameter estimates.[6] Whereas, in linear regression there is no such impact.

2. If instead there is a relationship between the observed response and the model predictors, independent of the true event risk – i.e., differential misclassification – the bias in the effect estimate can be positive or negative depending on the sign and magnitude of the relevant covariances.

In sum, misclassification in binary-outcome models not only suppresses true relationships – via attenuation or loss of power – but can also induce false positives through inflated effect estimates.[7]

To elaborate more formally, consider the familiar latent-variable representation of the binary outcome model

$$\mathbf{Y}^* = \beta_0 + \beta_1^t \mathbf{X} + \boldsymbol{\epsilon},$$

with latent-$\mathbf{Y}^*$ mapping onto the observed, censored outcome $\mathbf{Y_T}$ via the standard measurement equation

$$\mathbf{Y_T} = \mathbb{1}(\mathbf{Y}^* \geq 0),$$

where $\mathbf{Y_T}$ is the true outcome – equal to 1 if an event occurs. This has a probability of response

$$\text{pr}(\mathbf{Y_T} = 1|\mathbf{X}) = F(\beta_0 + \beta_1^t \mathbf{X}). \tag{1}$$

[6] In a toy simulation shows that with ≈ 25% misclassification the slope estimate in a logistic regression is less than half of the true value (0.4 vs. 1.0).

[7] Imai and Yamamoto (2010) discusses and evaluates the impact of differential measurement error on causal estimation in survey research, indicating that it can result in sizable overestimation of causal effects.

If, however, some set of outcomes in $\mathbf{Y_T} = 1$ is coded in $\mathbf{Y} = 0$, or vice versa, such that

$$\text{pr}(\mathbf{Y} = 1|\mathbf{Y_T} = 0) + \text{pr}(\mathbf{Y} = 0|\mathbf{Y_T} = 1) \neq 0,$$

there is misclassification. That is, misclassification occurs whenever the outcome vector, $\mathbf{Y}$, used in analysis, erroneously records some true events as zero and/or some non-events as one.[8]

With non-differential misclassification, the probability of accurate classification is

$$\begin{aligned} \text{pr}(\mathbf{Y} = 1|\mathbf{Y_T} = 1, \mathbf{X}) = \text{pr}(\mathbf{Y} = 1|\mathbf{Y_T} = 1) &= \pi_1, \\ \text{pr}(\mathbf{Y} = 0|\mathbf{Y_T} = 0, \mathbf{X}) = \text{pr}(\mathbf{Y} = 0|\mathbf{Y_T} = 0) &= \pi_0. \end{aligned}$$

This means that the probability for $\mathbf{Y}$ in not given by equation 1, but instead

$$\text{pr}(\mathbf{Y} = 1|\mathbf{X}) = (1 - \pi_0) + (\pi_1 + \pi_0 - 1)F(\beta_0 + \beta_1^t\mathbf{X}), \tag{2}$$

which equals equation 1 only if $\pi_0 = \pi_1 = 1$ (i.e., no misclassification).

As noted above, failing to account for misclassification results in inconsistent and biased effect estimates (Hausman, Abrevaya and Scott-Morton, 1998). Therefore, many strategies to address misclassification has received considerable attention elsewhere (Abrevaya and Hausman, 1999; Copas, 1988; Carroll and Pederson, 1993; Hausman, Abrevaya and Scott-Morton, 1998). Two problems persist with these existing remedies. First, many of these, including those enjoying the widest use currently in political science (Hausman, Abrevaya and Scott-Morton, 1998; Hug, 2009), simply maximize some version of the log-likelihood implied by equation 2. As noted by Carroll et al. (2006, p. 347), with these estimators "classification probabilities are only very weakly identified. . . parameters may be identified

[8] We argue that there are at least three causes of misclassification in social phenomena: *i)* misrepresentation, *ii)* misreporting, and *iii)* miscoding. In the first, agents under observation have the ability and an incentive to misrepresent its true type, behavior, or beliefs, and thus will supply inaccurate information. Secondly, misclassification can occur due to misreporting, wherein true information is revealed and available – e.g., an action occurs – but it is either not observed or properly recorded. Finally, misclassification can occur when an event is captured by a primary or secondary source, but miscoding errors occur in the construction of a data set from this, otherwise complete, information.

theoretically but not in any practical sense." As such, when and where possible, we will want to supply additional information to inform the misclassification probabilities over what we observe simply in $\mathbf{Y}$.

Second, none of these approaches is derived explicitly for the type of situation which motivates our project, that is, several sources of misclassified data. The Hausman, Abrevaya and Scott-Morton (1998) estimator, for example, is developed for a single misclassified binary outcome, not aggregate data from several sources erroneously treated as if it were a single binary outcome. As a result, the misapplication of this estimator to these data mismodels a fundamental feature of the data-generating process, resulting in a loss of efficiency or bias. In short, these estimators are designed to handle a different experimental condition from the one represented by the data considered heretofore. Therefore, we provide an alternative strategy in the next section.

3 A Multi-Source Solution

As discussed in the introduction, misclassification is likely to occur with media-based event data, where primary- or secondary-source reports fail to include the occurrence of an actual event. To introduce our method, consider two news outlets, 1 and 2, providing reports, $\mathbf{Y_1}$ and $\mathbf{Y_2}$, on the event of interest $\mathbf{Y_T}$.[9] Ultimately, we are interested in

$$\text{pr}(\mathbf{Y_T} = 1|\mathbf{X}),$$

where $\mathbf{X}$ is a matrix of predictors of the event. However, we possess two incomplete reports $\mathbf{Y_1} \neq \mathbf{Y_T}$ and $\mathbf{Y_2} \neq \mathbf{Y_T}$, explained by:

$$\text{pr}(\mathbf{Y_1} = 1|\mathbf{X}, \mathbf{Z_1})$$

$$\text{pr}(\mathbf{Y_2} = 1|\mathbf{X}, \mathbf{Z_2}),$$

[9] While we focus on the two-source model during elaboration and evaluation, it is easy to extend this to accommodate additional sources as we show in Section 6.

where $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are predictors of the (mis-)reporting of an event (e.g., distance to reporting office) by that source, which are otherwise unrelated to $\mathbf{Y_T}$. Following convention in the applied literature, we aggregate these sources to reduce the individual missingness by

$$\mathbf{Y_{sum}} = \mathbb{1}(\mathbf{Y_1} + \mathbf{Y_2} \geq 1).$$

If $\mathbf{Y_{sum}} = \mathbf{Y_T}$, the data are complete and we find that

$$\text{pr}(\mathbf{Y_T} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = \text{pr}(\mathbf{Y_T} = 1|\mathbf{X}) = F(\beta^t \mathbf{X}), \tag{3}$$

where $F(\cdot)$ is specified up to the parameter β. However, where $\mathbf{Y_{sum}} \neq \mathbf{Y_T}$, we are unable to simplify as in 3. This means that when observed outcomes are misclassified, fitting Equation 3 will result in biased estimates of $\mathbf{X}$ on $\mathbf{Y_T}$.

Therefore, we construct an estimator around

Assumption 1 We make the following assumptions: (a) $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are independent given $(\mathbf{Y_T}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$. (b) $\mathbf{Y_{sum}} = 1$ implies $\mathbf{Y_T} = 1$ with probability 1. (c) $\mathbf{Y_T} = 0$ implies that $\mathbf{Y_{sum}} = 0 = \mathbf{Y_1} = \mathbf{Y_2}$ with probability 1.

Less formally, Assumption 1(a) states that the sources of reporting data are conditionally independent of one another. Assumption 1(b) and Assumption 1(c) jointly indicate that misclassification in this context is exclusively underreporting.

If we treat $\mathbf{Y_{sum}}$ as the response variable, the problem is related to one studied by Copas (1988), Carroll and Pederson (1993) and Hausman, Abrevaya and Scott-Morton (1998). The misclassification probabilities are

$$\begin{aligned} \text{pr}(\mathbf{Y_{sum}} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}), & (4) \\ \text{pr}(\mathbf{Y_{sum}} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= 0, & (5) \end{aligned}$$

where (5) followed from Assumption 1(b). In Hausman, Abrevaya and Scott-Morton (1998) these misclassification probabilities do not depend of the covariates, and are instead simply an unknown constant to be estimated.[10] Therefore, we generalize Hausman, Abrevaya and Scott-Morton (1998)'s estimator to allow for misclassification probabilities that are dependent upon the covariates (as shown in **??**)

$$\begin{aligned}
\mathrm{pr}(\mathbf{Y_{sum}} = 0|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})\}\{1 - F(\mathbf{X}, \beta)\} + \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}); \qquad (6)\\
\mathrm{pr}(\mathbf{Y_{sum}} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})\}F(\mathbf{X}, \beta). \qquad (7)
\end{aligned}$$

In principle, since the form of $F(\cdot)$ is assumed known, then $\gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$ is identified nonparametrically. If one assumes a parametric form for $\gamma(\cdot)$, then maximum likelihood can be employed.

However, the data are not $(\mathbf{Y_{sum}}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$, but $(\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$, that is, we have multiple sources of data. As such, there may be different misclassification rates, which is the fundamental difference between our estimator and existing approaches. Rather than neglect this information, thereby failing to use all of the data, we define

$$\begin{aligned}
\alpha_1(\mathbf{X}, \mathbf{Z_1}) &= \mathrm{pr}(\mathbf{Y_1} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_1}); \qquad (8)\\
\alpha_2(\mathbf{X}, \mathbf{Z_2}) &= \mathrm{pr}(\mathbf{Y_2} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_2}). \qquad (9)
\end{aligned}$$

Here by Assumption 1(b) we have that $\mathrm{pr}(\mathbf{Y_1} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_1}) = \mathrm{pr}(\mathbf{Y_2} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_2}) = 0$. Then under Assumption 1(a), it is easy to see that

$$\gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = \alpha_1(\mathbf{X}, \mathbf{Z_1})\alpha_2(\mathbf{X}, \mathbf{Z_2}). \qquad (10)$$

Indeed, in Online Appendix A, we show the following result.

Lemma 1 Under Assumption 1,

$$\mathrm{pr}(\mathbf{Y_1} = 0, \mathbf{Y_2} = 0|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = 1 - F(\mathbf{X}, \beta) + \alpha_1(\mathbf{X}, \mathbf{Z_1})\alpha_2(\mathbf{X}, \mathbf{Z_2})F(\mathbf{X}, \beta);$$

[10] Hausman, Abrevaya and Scott-Morton (1998) allude a generalization of their estimator which would permit the inclusion of exogenous predictors of the misclassification probabilities, though they never return to fully elaborate on such an approach. In a follow up work, Abrevaya and Hausman (1999) do devote greater attention to covariate-dependent measurement error in a semiparameteric framework.

$$
\begin{aligned}
\mathrm{pr}(\mathbf{Y_1} = 0, \mathbf{Y_2} = 1 | \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \alpha_1(\mathbf{X}, \mathbf{Z_1})\{1 - \alpha_2(\mathbf{X}, \mathbf{Z_2})\}F(\mathbf{X}, \beta); \\
\mathrm{pr}(\mathbf{Y_1} = 1, \mathbf{Y_2} = 0 | \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \alpha_1(\mathbf{X}, \mathbf{Z_1})\}\alpha_2(\mathbf{X}, \mathbf{Z_2})F(\mathbf{X}, \beta); \\
\mathrm{pr}(\mathbf{Y_1} = 1, \mathbf{Y_2} = 1 | \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \alpha_1(\mathbf{X}, \mathbf{Z_1})\}\{1 - \alpha_2(\mathbf{X}, \mathbf{Z_2})\}F(\mathbf{X}, \beta).
\end{aligned}
$$

This implies the likelihood function

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

Under our assumptions, this allows us to estimate parameters for the risk model and misclassification probabilities. This improves over current estimators which either force researchers to erroneously assume that there is no misclassification in the data, that the data originates from one source, or both. Furthermore, our estimator utilizes more data-based information to achieve identification, resulting in sounder estimates of the misclassification and, in turn, event probabilities.[11]

In sum, our estimator allows researchers using event-based data to accurately test theories on both the event of interest and (mis-)reporting of these events.[12]

[11] To present some intuition non-technically, consider the canonical mark-recapture example of attempting to estimate the number of fish in a pond. If we cast a net only once, the only data-based information we have is how many fish are in the net. Considering regression models, naïve probit treats this is the complete population of fish in the pond (e.g., no misclassification), whereas the Hausman, Abrevaya and Scott-Morton (1998) approach attempts to guess how many fish remain in the pond given how many we've captured in the net - both are flawed. Instead, as anyone familiar with mark-recapture has surmised, we want to 'mark' the first catch, release them, and cast the net a second time. Now the number of those fish re-captured can be used to generate estimates on the total number of uncaught fish remaining in the pond. This, as we see it, is our estimator, where we can use whether only one or multiple sources reported on some event to generate accurate estimates. See Hendrix and Salehyan (2015) for a discussion on considering event-based data as a mark-recapture problem.

[12] Note that the ability of our model to recover accurate estimates of the event probability depends on obtaining accurate estimates of the misclassification probabilities, that is, in properly specifying the misclassification model. As such, researchers should consider this specification with the same care they devote to modeling the event itself.

4 Simulations

Our simulation study is designed to evaluate the performance of estimators under outcome misclassification. We consider the following five methods:

1. **Naïve Probit**: fits a probit model to Y_{sum} with event probability $\text{pr}(Y_{\text{sum}} = 1|X) = \Phi(\beta_0 + \beta_1 X)$.

2. **Hausman, Constant Probabilities**: the approach outlined in Hausman, Abrevaya and Scott-Morton (1998) – assuming constant misclassification probabilities and exploiting our assumption 1(b)[13] – which fits $\text{pr}(Y_{\text{sum}} = 1|X) = \{1 - \pi_1\}\Phi(\beta_0 + \beta_1 X)$.

3. **Hausman with Covariates**: our generalization of Hausman, Abrevaya and Scott-Morton (1998) allowing for non-constant misclassification probabilities $\text{pr}(Y_{\text{sum}} = 0|Y_T = 1, X, Z_1, Z_2) = \gamma(X, Z_1, Z_2) = \Phi(\eta_{00} + \eta_{01}X + \eta_{02}Z_1 + \eta_{03}Z_2)$, giving event probabilities $\text{pr}(Y_{\text{sum}} = 1|X, Z_1, Z_2) = \{1 - \gamma(X, Z_1, Z_2)\}\Phi(\beta_0 + \beta_1 X)$.

4. **Multi-Source, Constant Probabilities**: our multi-source method detailed in Section 3, but restricted to use constant probabilities $\alpha_1(X, Z_1) = \Phi(\eta_{10})$ and $\alpha_2(X, Z_2) = \Phi(\eta_{20})$.

5. **Multi-Source with Covariates**: our general multi-source method detailed in Section 3 with $\alpha_1(X, Z_1) = \Phi(\eta_{10} + \eta_{11}X + \eta_{12}Z_1)$ and $\alpha_2(X, Z_2) = \Phi(\eta_{20} + \eta_{21}X + \eta_{22}Z_2)$.

4.1 Simulation Design

The data-generating process for the simulations is the following:

- Step I. Take n draws of $X, Z_1$and Z_2 from a $N(0, 1)$ distribution.

- Step II. Generate Y_T from a Bernoulli distribution with success probability $F(X_i, \beta) = \Phi(\beta_0 + X_i\beta_1)$, $i = 1, \ldots, n$, where Φ denotes the CDF of the standard normal distribution.

[13] In general the Hausman, Abrevaya and Scott-Morton (1998) estimator does not require 1(b), allowing for $\pi_0 \neq 0$. We evaluated this method as well, however, found that it failed to converge approximately 80% of the time under our simulated conditions. As such, we do not report these results.

- Step III. Generate misclassification probabilities using

$$\begin{aligned}\alpha_1(X_i, Z_{i,1}) &= \Phi(\eta_{10} + \eta_{11}X_i + \eta_{12}Z_{i,1}), \text{ and} \\ \alpha_2(X_i, Z_{i,2}) &= \Phi(\eta_{20} + \eta_{21}X_i + \eta_{22}Z_{i,2}),\end{aligned}$$

  then generate $Y_{i,1}$ and $Y_{i,2}$

$$Y_{i,1} = Y_{i,T}(1 - B(\alpha_1)) \text{ and } Y_{i,2} = Y_{i,T}(1 - B(\alpha_2)).$$

- Step IV. Given Y_1 and Y_2, generate Y_{sum} using $Y_{\text{sum}} = \mathbb{1}(Y_{i,1} + Y_{i,2} \geq 1)$.

Across all experiments, we generate $N = 1000$ data sets (i.e., trials), each with sample size $n = 1000$. Our experimental data-generation process nests all the methods detailed above, with different assignments to the η's producing each of these as the true model. As such, we are mainly interested in the effect of varying those parameters, so we fix $\beta_0 = -1$ and $\beta_1 = 1$ in all experiments (producing $\text{pr}(Y_{\text{sum}} = 1) \approx 0.30$).

We investigate the effect of misclassification under two broad sets of experimental conditions produced from different specifications in Step III:

1. Non-differential misclassification – the misclassification errors α_1 and α_2 do not depend on the covariates and are constant (i.e., $\eta_{10} > 0$, $\eta_{20} > 0$, and $\eta_{11} = \eta_{12} = \eta_{21} = \eta_{22} = 0$)[14]

2. Differential misclassification – the misclassification errors α_1 and α_2 depend on the event covariates (i.e., $\eta_{11} > 0$ and/or $\eta_{21} > 0$)

In the next section we provide detailed results from two experiments under these conditions. In Experiment 1 we set $\alpha_1 = 0.35$ and $\alpha_2 = 0.2$,[15] Drawing on previous studies which have evaluated the extent of misclassification in data of these types, we consider this a moderate level of misclassification which is likely to be observed by researchers. In Experiment 2,

[14] Note that non-differential misclassification only requires that $\eta_{11} = \eta_{21} = 0$, with η_{12} and η_{22} determining whether these probabilities also vary across units.

[15] Continuing with our notation above this is equivalent to setting η_{10} to -0.3885 and η_{20} to -0.841, with all other η's at zero.

we set $(\eta_{10}, \eta_{11}, \eta_{12}, \eta_{20}, \eta_{21}, \eta_{22}) = (-0.7, 1, 1, -1.4, 1, 1)$, so that $E\{\alpha_1(X, Z_1)\} \approx 0.35$ and $E\{\alpha_2(X, Z_2)\} \approx 0.2$. This allows us to evaluate distinct effects of non-differential and differential error under roughly the same rate of misclassification.[16]

4.2 Results

The results of the simulation study are presented in Table 1, with the bias, standard deviation (STD), estimated standard error (SE), and mean squared error (MSE), and 95% coverage probability (CP) reported.[17] The top half, Experiment 1, presents the results of our constant, non-differential error simulations. We see that, as expected, our *Multi-Source* methods (Models 4 & 5) outperform the other methods in mean-square error terms. Furthermore, the coverage probabilities – the proportion of simulations in which the parameter is contained in the interval estimate – for both closely reflect the nominal 95% confidence levels. The alternative estimators, on the other hand, perform noticeably worse. The *Hausman with Covariates* (Model 3), is obviously flawed when the misclassification probabilities in the DGP are fixed, with estimates varying wildly from simulation to simulation. The *Naïve Probit* and *Hausman Constant* estimators perform better than this, but still underperform our proposed estimators. *Naïve Probit* (Model 1) does only slightly worse in mean-square error terms, however, the bias and resultant anti-conservative coverage probabilities are troubling, while the *Hausman*-type estimator (Model 2) is nearly two-times worse than our estimators in MSE.

While our estimators do perform well, we do not want to overstate the extent of the gains. In general, the conventional estimation strategies seem to do fairly well if the misclassification rates are truly non-differential. However, this rarely happens in practice as underreporting is usually systemic, that is, there is a reason why some observations are misclassified and not others (Schrodt, 2012).[18] Experiment II, the lower half of Table 1, provides the results

[16] All simulations were completed in R. The code for our estimator, *Multi-Source with Covariates*, is provided in Online Appendix B and code for all novel estimations strategies presented – i.e., Methods 3, 4, and 5 – will be made available for public use.

[17] All replication materials can be found online at Cook et al. (2016).

[18] An analogous problem for missing data may be more familiar to our readers, where the related distinction would be between data missing completely at random (MCAR) and missing not at random (MNAR). When data are MNAR researchers require a model predicting the missingness in their data, as we need a model

from simulations under these conditions. We observe a substantial degradation in the performance of the conventional strategies. The *Naïve Probit* and *Hausman Constant* estimators have substantial bias in the slope estimates, with MSE orders of magnitude larger than our preferred methods. Moreover, the *Naïve Probit* estimator *never*(!) recovers the true sample statistic in any of the simulations (CP = 0.0). As this is the dominant empirical strategy used in political science this is clearly a problem.[19] We see expected gains in the *Hausman with Covariates* (Model 3) and degradation in *Multi-Source Constant*, reflecting the accuracy with which they capture the true data-generating process. Our *Multi-Source with Covariates* method clearly dominates, with the lowest MSE and most accurate coverage probabilities. More importantly, perhaps, is that our *Multi-Source with Covariates* method is robust to either type of misclassification – i.e., differential or non-differential – as it is nearly dominant in both sets of simulations. Thus, researchers can employ this method when they do not have strong ex ante beliefs over the cause of misclassification in their data and be confident in the results obtained.

Parameter estimates are often not directly the quantity of interest. Instead, researchers are interested in some transformation of the parameter, such as the marginal effect, which is not equal to the reported coefficient in all but the linear-additive model. As such, we calculate the marginal effect of X for each of the estimators as

$$\partial Y/\partial X = \Phi\{\hat{\beta}_0 + \hat{\beta}_1(\mu_x + \sigma_x)\} - \Phi\{\hat{\beta}_0 + \hat{\beta}_1\mu_x\}.$$

The results of both experiments are given in Table 2. For Experiment 1, we see that as in parameter estimation, all estimators perform quite well in terms of MSE. *Naïve Probit* has the downward bias we would expect from attenuation in non-differential misclassification, yet the results from all estimators are fairly encouraging. Experiment II is quite different, here we see substantial bias in each of the estimators that fails to model the misclassification probabilities as a function of covariates. Interestingly, we see a large attenuating bias in *Naïve*

predicting misclassification here.

[19] The shortcomings we evidence here would also occur in a logistic regression, the more important consideration is not the link function but whether the estimator accounts for misclassification. Note that our estimator is easily extended to allow a logistic functional form, we merely use the probit for easier comparison to the Hausman estimator and evaluation of correlated outcomes.

Table 1: Simulation Study Results

		Method				
		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-Source	Multi-Source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
		Experiment 1: $\alpha_1 = 0.35, \alpha_2 = 0.2$				
$\beta_0 = -1$	Bias	0.052	-0.007	-1.674	0.004	0.002
	STD	0.058	0.092	8.449	0.063	0.064
	SE	0.060	0.092	3.390	0.063	0.064
	MSE	0.006	0.009	74.111	0.004	0.004
	CP(%)	87.5	92.2	95.8	95.9	95.7
$\beta_1 = 1$	Bias	0.054	-0.028	-1.368	-0.007	-0.005
	STD	0.066	0.111	9.400	0.075	0.078
	SE	0.067	0.110	3.022	0.077	0.079
	MSE	0.007	0.013	90.137	0.006	0.006
	CP(%)	85.7	93.2	93.0	95.5	95.3
		Experiment 2: $\alpha_1 = \Phi(-0.7 + X + Z_1)$, $\alpha_2 = \Phi(-1.4 + X + Z_2)$				
$\beta_0 = -1$	Bias	0.074	-0.663	-0.023	-0.048	0.003
	Std	0.051	0.227	0.106	0.060	0.063
	SE	0.055	0.175	0.096	0.067	0.063
	MSE	0.008	0.490	0.012	0.006	0.004
	CP(%)	77.4	3.7	95.8	91.3	95.8
$\beta_1 = 1$	Bias	0.397	-0.379	-0.035	0.287	-0.010
	Std	0.055	0.258	0.158	0.092	0.094
	SE	0.057	0.218	0.132	0.079	0.095
	MSE	0.161	0.210	0.026	0.091	0.009
	CP(%)	0.0	56.9	94.0	11.1	96.3

Note: Included in the table is the average bias of the estimator (Bias), the standard deviation of the estimates across the simulations (STD), the mean estimated standard error (SE), the mean-squared error (MSE), and the coverage probabilities of a nominal 95% confidence interval (CP). The estimation methods and experimental conditions are detailed in Section 4.

Probit and a large inflationary bias in *Hausman Constant*, suggesting that rather than solve the problem the *Hausman*-type estimator simply introduces a new one. As before, we see that our preferred method, *Multi-Source with Covariates*, provides accurate marginal effect estimates under other experimental condition.

Table 2: Marginal Effects in Simulation Studies

		Method				
		(1)	(2)	(3)	(4)	(5)
		Naive	Hausman	Hausman	Multi-Source	Multi-Source
Parameter		Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
				Experiment 1: $\alpha_1 = 0.35, \alpha_2 = 0.2$		
$\partial Y/\partial X$	Bias	-0.030	0.011	-0.038	0.002	0.001
	Std	0.023	0.049	0.109	0.027	0.028
	MSE	0.001	0.002	0.013	0.001	0.001
				Experiment 2: $\alpha_1 = \Phi(-0.7 + X + Z_1)$, $\alpha_2 = \Phi(-1.4 + X + Z_2)$		
$\partial Y/\partial X$	Bias	-0.164	0.119	0.012	-0.106	0.003
	Std	0.019	0.070	0.059	0.038	0.036
	MSE	0.027	0.019	0.004	0.013	0.001

Note: Included in the table is the average bias of the estimator (Bias), the standard deviation of the estimates across the simulations (STD), and the mean-squared error (MSE). The estimation methods and experimental conditions are detailed in Section 4.

While our estimator performs well under either type of misclassification, two additional issues may be of concern with real data. First, researchers will often not have complete models of misclassification, as such it will be important to know the affect of omitted variables in the misclassification sub-models. Second, often sources will not perfectly reflect Assumption 1(a) – i.e., local independence across sources – which was part of the derivation of our estimator above, as such it will be important to know the small-sample implications of violations of this condition.[20] We explore both concerns in an additional series of simulations where we include additional, correlated unobservables in the generation of the misclassification

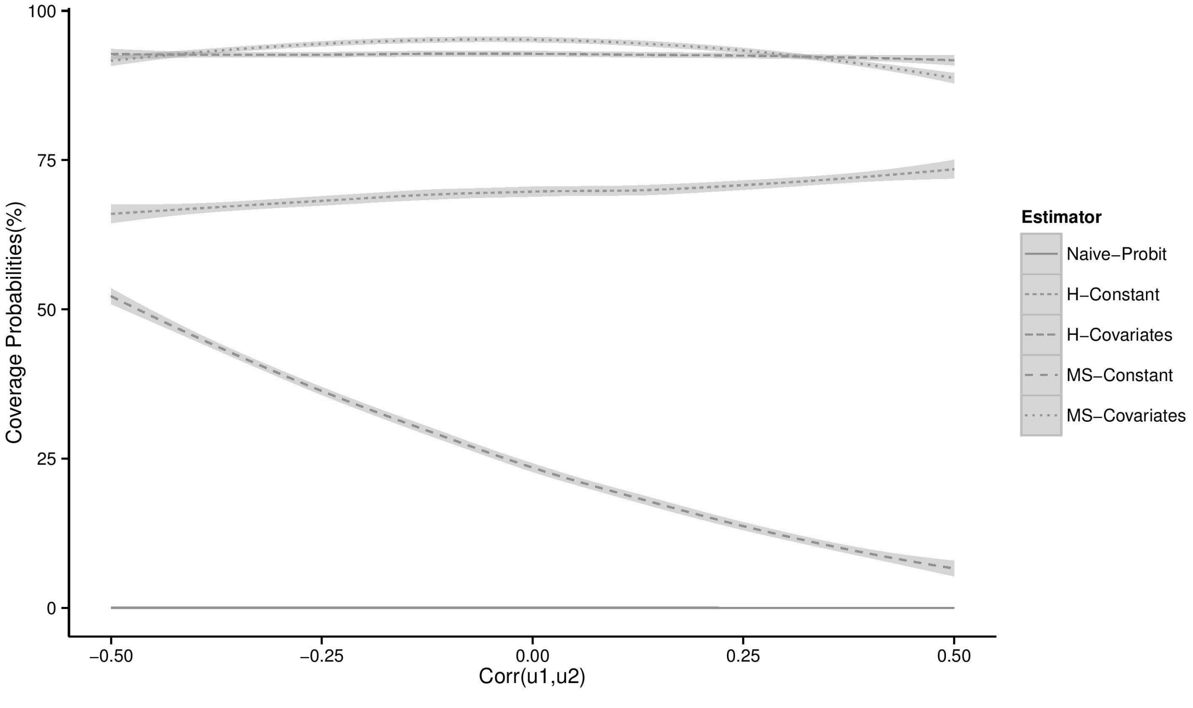
[20] Strict identification of the parameters, however, does not depend on the maintenance of Assumption 1(a). In extensions we show that identification with parametric models can be achieved though just assumptions 1(b) and 1(c).

probabilities:

$$\begin{aligned}\alpha(X_i, Z_{i,1}) &= \Phi(\eta_{10} + \eta_{11}X_i + \eta_{12}Z_{i,1} + u_1),\\ \alpha(X_i, Z_{i,2}) &= \Phi(\eta_{20} + \eta_{21}X_i + \eta_{22}Z_{i,2} + u_2),\end{aligned}$$

where $(u_1, u_2)^{\mathrm{T}} = \mathrm{Normal}(0, \Sigma)$, where Σ is the covariance matrix of a bivariate standard normal random variable with variance = 1 and correlation ρ, with changes to ρ varying the extent of the correlation across the sources. We run 51 additional simulations for Experiment 2 with this modification – evaluating correlations from $\rho = -0.5$ to $\rho = 0.5$ in increments of 0.02 – capturing both strong negative correlation (i.e., sources purposefully report on distinct events) and strong positive correlation (i.e., sources purposefully report on the same events). The results are illustrated in Figure 1, which provides the coverage probabilities across these simulations.

Figure 1: Estimator Performance with Source Correlation



Note: Curves smoothed via LOESS. Points or non-smoothed lines render the same conclusions.

We see that our preferred method, *Multi-Source with Covariates*, consistently performs well even despite omitted variables in the misclassification equations and high levels of correlation. With correlation between -0.44 and 0.3 it is the optimal estimator, bested only by the *Hausman with Covariates* when the level of correlation becomes very high. This is expected given that under these conditions, extremely high levels of correlation, there is less additional information from the second source – they converge toward one another. What is more notable is the persistent fitness of our estimator across all simulations, indicating that it is robust to typical violations of Assumption 1(a) one might observe in real-world data like that we consider in the next section.[21]

5 State Repression in Africa

As discussed in the introduction, event-based data have been widely used in the literature on contentious politics. Misclassification is widely believed to be an issue in these studies, with researchers arguing that such "incorrect misclassification is likely to be systemic rather than random." (Schrodt, 2012, 557) That is, outcome data is likely to suffer from the type of error observed to generate substantial bias in our simulations. Notably, the literature on repression has generated theories predicting when and where we should be most likely to observe underreporting; arguing that high-visibility events occurring in urban centers of economically developed, less authoritarian regimes are more likely to be reported (Davenport, 2007; Davenport and Ball, 2002). Yet, much of the quantitative literature on repression fails to explicitly account for the potential bias introduced from underreporting in their statistical analysis. This motivates our current analysis, where we examine both: *i*) the effect of misclassification on common predictors of repression and *ii*) analyze whether we find support for those factors thought to produce reporting bias.

Specifically, we estimate a model of repression in Africa using the set of methods detailed in Section 4. Our outcome data is taken from SCAD (Salehyan et al., 2012), which generates event data on forty-seven African countries using key word searches of Associated Press (AP) and Agence France-Presse (AFP) news wires. These data are particularly useful for

[21] A range of additional simulated conditions were explored including: mixed misclassification, omitted variables in the misclassification model, multiple predictors in the risk model, equal probabilities of misclassification across sources. We describe, present, and discuss these additional experiments in Online Appendix C.

our purposes in two ways. First, since 2012 events have been recorded as being reported by the AP, the AFP, or both – that is, there are multiple sources.[22] Second, the creators of this data set have discussed the likelihood of underreporting, utilizing a Lincoln-Peterson mark-recapture method to estimate that 24% of social conflict events go unreported (Hendrix and Salehyan, 2015).[23] While mark and recapture methods prove useful for diagnosing the presence of underreporting, they do not remedy the resulting bias in subsequent empirical analysis as we aim to do.[24]

Therefore, using these data, we generate a binary outcome, *Repression*, which is coded as one if either or both of the news wires report on (lethal or non-lethal) repression initiated by the government or pro-government actors during a state-month, and zero otherwise. For the sub models in our misclassification estimator, we construct two additional binary variables, one for events reported by the AP and one for events reported by the AFP. Following the Poe and Tate (1994) model, we believe repression should be increasing in population, and decreasing in democracy and GDP per capita.[25] We include each as independent variables, with coding and data sources elaborated in the Online Appendix. Lastly, our estimator requires additional covariates predicting misclassification, which is unrelated to the true event probability. Hendrix and Salehyan (2016) have suggested that non-conflict related news reports may indicate the total amount of media effort devoted to a country and include a country-year average (divided into quantiles) as a control in their model. Drawing on this, we collect new data on the number of non-conflict related news reports for each country by the AP and AFP respectively.[26] The values for each, *AP Reports* and *AFP Reports*, are introduced in the sub-models predicting misclassification. Our expectation is that greater media effort (e.g., higher values) will be negatively associated with the probability of misclassification.

The results from these models are presented in Tables 3 and 4. Table 3 shows the results from the repression model of each estimator, the outcome of primary theoretical interest in

[22] Prior to this the SCAD data simply indicate whether or not there were multiple sources.

[23] Using the LP estimator, we calculate that 14% of repression events go unreported.

[24] If one assumed constant misclassification probabilities the singular estimate of misclassification from mark-recapture methods could be built into a weighted likelihood, however, when the risk of misclassification is a function of time-varying covariates such an approach is infeasible.

[25] Elaborated and clarified in Poe, Tate and Keith (1999); Poe, Rost and Carey (2006)

[26] Specifically, we used the Boolean opposites of the SCAD search terms – protest, riot, strike, violence, attack – and counted the number of (non-violent) news stories.

Table 3: Model of Repression in Africa

	(1)	(2)	(3)	(4)	(5)
	Naive	Hausman	Hausman	Multi-Source	Multi-Source
Model	Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
$GDPpc_{t-1}$	0.020	0.020	−0.164	0.022	−0.292
	(0.062)	(0.062)	(0.125)	(0.072)	(0.145)
Pop_{t-1}	0.407	0.407	0.314	0.458	0.330
	(0.053)	(0.053)	(0.085)	(0.063)	(0.095)
$Demo_{t-1}$	−0.655	−0.655	−0.739	−0.757	−0.819
	(0.151)	(0.151)	(0.261)	(0.172)	(0.315)
Constant	−8.011	−8.011	−4.679	−8.568	−3.857
	(1.000)	(0.994)	(1.624)	(1.161)	(2.063)
N	1092	1092	1092	1092	1092

Note: The estimators are the same as those used in the simulations and detailed in Section 4. The covariates are the log of GDP per capita (*GDPpc*), the log of population (*Pop*), and an indicator if the country is a democracy (*Demo*), each lagged by one period.

our illustration. Glancing across the table highlights both *i*) the importance of accounting for misclassification and *ii*) how one accounts for misclassification. With the *Naïve Probit* (Model 1), we see that both *Pop* and *Demo* are significant in their expected directions, with *GDPpc* positive and insignificant. The effect of *GDPpc* is inconsistent with the theoretical literature, however, it is not uncommon in the empirical literature to date.[27] We see that the results from the Hausman, Abrevaya and Scott-Morton (1998) estimator (Model 2) are nearly identical to those from the naïve probit in this analysis. This, despite the fact that theory would suggest, and the originators of the data set have concluded, that events go underreported in the data, that is, exactly the setting in which researchers would turn to this estimator.[28] The *Multi-Source Constant* model also does not affect much change, with all results roughly the same as in Models 1 and 2.

[27] Hendrix and Salehyan (2016), with a wider sample and additional predictors even finding an unexpected positive and significant effect of the log of GDP per capita.

[28] We believe this to be an artifact of the numerical instability of the Hausman-type estimator, which has been found elsewhere before (Hug, 2009).

None of this is surprising given that our belief is that the misclassification is systemic. As such, we turn to Models 3 and 5 where the misclassification probabilities are non-constant, predicted by the same covariates included in the repression model. We see sizable differences in Model 3, *Hausman with Covariates*, with an increase in the constant offset by decreases in both *Pop* and *Demo*. That is, the base-line risk of repression is more likely than what is evidenced in our reported sample due to misclassification, which also appears to have biased the effect of the predictors. Similar, if more pronounced, results are found in Model 5, our *Multi-Source with Covariates* model, with fairly dramatic shifts to all 3 predictors and the constant. Most notably, *GDPpc* becomes negative and is now statistically significant, consistent with theoretical expectations. We also see the negative effect of *Demo* increases and the positive effect of *Pop* increases. What does this mean? It suggests misclassification was biasing the effect of *GDPpc*, *Pop*, *Demo* as repression in wealthy, populous democracies is more likely to be reported on, causing us to erroneously conclude that repression is actually more likely in those environments than it is.[29]

We can examine these causes of reporting bias more explicitly in Table 4, which provides the misclassification probabilities (and models) where they are estimated.[30] Focusing on Model 5, we observe two main findings of note: first, the AP is more likely to suffer from underreporting than AFP, as indicated by their constants; second, our measures for "media effort" are significantly and negatively related to misclassification, that is, the more non-violent news stories reported on a region by the AP and AFP, the more likely they are to accurately report an incident of repression. We do not find significant support for the repression-model predictors in this analysis, however, given that these should predict media effort, not just misclassification, this null finding makes sense as we already account for

[29] Our contention is not that this is a perfect theoretical model of state repression. We readily admit its limitations as a more general model of repression, as we are constrained (due to temporal coverage) from including several additional predictors of repression that one would commonly find in the literature. However, the confined focus of our spatial sample (i.e., African countries) of our analysis helps reduce the need for extraneous covariates to gain balance (Achen, 2002). Additionally, the main purpose of our analysis is, first and foremost, to illustrate the extent to which estimates are sensitive to misclassification in the dependent variable. While additional covariates may alter some of the parameter estimates, it would not change this fundamental reality as it would not cause outcomes currently contaminated to become correctly classified.

[30] Note that the constant in model 2 refers to the constant misclassification probability estimate for both sources, not simply the AP responses as the table layout may suggest. That is, the Hausman estimator here only provides a single such estimate given that it does not account for the multi-source nature of the data.

Table 4: Models of Reporting Bias on Repression in Africa

	(1)	(2)	(3)	(4)	(5)
	Naive	Hausman	Hausman	Multi-Source	Multi-Source
	Probit	Const Pr	w/ Cov	Const Pr	w/ Cov
			Misclassification model for AP		
GDPpc			−0.083		−0.280
			(0.259)		(0.168)
Population			−0.006		−0.268
			(0.154)		(0.106)
Democracy			0.229		−0.386
			(0.508)		(0.416)
AP Reports			0.018		−0.033
			(0.021)		(0.008)
Constant		0.001	1.946	0.559	7.947
		(0.003)	(3.076)	(0.148)	(2.073)
			Misclassification model for AFP		
GDPpc					−0.203
					(0.172)
Population					−0.057
					(0.121)
Democracy					0.350
					(0.402)
AFP Reports			−0.086		−0.023
			(0.025)		(0.006)
Constant				0.005	3.332
				(0.181)	(2.359)
N	1092	1092	1092	1092	1092

Note: Results produced by the same models estimated in Table 3, partitioned for ease of exposition. *AP Reports* & *AFP Reports* are number of non-conflict news stories. To clarify the presentation: Model (1) produces no estimates of misclassification; Model (2) estimates a single misclassification probability, common to both sources; Model (3) estimates a misclassification model of the dependent variable, from either source, using the risk model covariates and *AP Reports* & *AFP Reports*; Model (4) estimates separate, constant misclassification probabilities for the two sources; and Model (5) estimates separate misclassification models for the two sources.

media effort explicitly as an additional predictor.

6 Applications & Extensions

To introduce and describe our approach we have focused our discussion on fixed observational units with only two binary-event indicators (e.g., reports) of a binary outcome, however, particular applications of concern to applied researchers may deviate from this in several ways. Therefore, we detail some of the more likely departures here and discuss how our method can be utilized under a variety of these contexts.

First, data are often compiled from more than two underlying sources. As alluded to in footnote 9, our method can be easily amended to handle these additional sources by simply expanding the joint likelihood. In the most general setting, suppose that there are $M(\geq 2)$ reporting sources. Let the binary outcome variable from the j^{th} source be $\boldsymbol{Y}_j$, $j = 1, \ldots, M$, and $\boldsymbol{Y}_{\mathrm{T}}$ be the true indicator of an event. Let $\boldsymbol{X}$ be a covariate that is associated with the true event indicator, and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M$ be the source specific covarites. Maintaining all earlier assumptions, define source specific false negative probabilities, $\alpha_j(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathrm{pr}(\boldsymbol{Y}_j = 0|\boldsymbol{Y}_{\mathrm{T}} = 1, \boldsymbol{X}, \boldsymbol{Z}_j)$, for $j = 1, \ldots, M$, and $\mathrm{pr}(\boldsymbol{Y}_{\mathrm{T}} = 1|\boldsymbol{X}) = F(\boldsymbol{X}, \beta)$ Then the joint probability of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M$ given $\boldsymbol{X}$, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M$ is

$$\begin{aligned}
& \mathrm{pr}(\boldsymbol{Y}_1 = y_1, \ldots, \boldsymbol{Y}_M = y_M|\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M) \\
= \; & \mathrm{pr}(\boldsymbol{Y}_1 = y_1, \ldots, \boldsymbol{Y}_M = y_M|\boldsymbol{Y}_{\mathrm{T}} = 0, \boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)\mathrm{pr}(\boldsymbol{Y}_{\mathrm{T}} = 0|\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M) \\
& + \mathrm{pr}(\boldsymbol{Y}_1 = y_1, \ldots, \boldsymbol{Y}_M = y_M|\boldsymbol{Y}_{\mathrm{T}} = 1, \boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)\mathrm{pr}(\boldsymbol{Y}_{\mathrm{T}} = 1|\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M) \\
= \; & \{1 - F(X, \beta)\}I(y_1 = \cdots = y_M = 0) \\
& + \alpha_1^{1-y_1}(\boldsymbol{X}, \boldsymbol{Z}_1)\{1 - \alpha_1(\boldsymbol{X}, \boldsymbol{Z}_1)\}^{y_1} \times \cdots \times \alpha_M^{1-y_M}(\boldsymbol{X}, \boldsymbol{Z}_M)\{1 - \alpha_M(\boldsymbol{X}, \boldsymbol{Z}_M)\}^{y_M} F(\boldsymbol{X}, \beta),
\end{aligned}$$

for $y_j = \{0, 1\}$ and $j = 1, \ldots, M$. Therefore, the likehood function of the observed data over N independent observations is

$$L = \prod \Bigg[\{1 - F(\boldsymbol{X}, \beta)\}I(y_1 = \cdots = y_M = 0)$$

$$+\alpha_1^{1-y_1}(\boldsymbol{X}, \boldsymbol{Z}_1)\{1-\alpha_1(\boldsymbol{X}, \boldsymbol{Z}_1)\}^{y_1} \times \cdots \times \alpha_M^{1-y_M}(\boldsymbol{X}, \boldsymbol{Z}_M)\{1-\alpha_M(\boldsymbol{X}, \boldsymbol{Z}_M)\}^{y_M} F(\boldsymbol{X}, \beta)\Bigg],$$

where the outer product is over N. As always, the maximum likelihood estimator of the model parameters can be obtained by maximizing the logarithm of L.[31]

This shows how our estimation strategy can easily be tailored for a general number of sources (e.g., 3, 5, 8...) in a straightforward way. However, some event data sets are compiled from hundreds or even thousands of sources (e.g., ICEWS, GDELT, Phoenix, etc.), which would tax our method and result in a badly behaved likelihood; as the information in each 'cell' of the joint likelihood becomes increasingly small. A partial solution for these data may be to classify the many sources into fewer clusters (e.g., international news agencies, national newspapers, local newspapers) which share common, within-cluster, features in the scope and nature of their coverage. Then it is reasonable to assume that the misclassification probabilities vary across the clusters but not much within a cluster, reducing the number of parameters and permitting estimation as before.

As an example, consider the Armed Conflict Location & Event Dataset (ACLED), a widely-used dataset on sub-state violent events.[32] These data are drawn from a variety of sources – ranging between 70 and 232 sources from 1997 to 2012 – which, at first pass, might suggest that our approach is not feasible. However, the top-10 sources account for nearly three-fourths of the events in the dataset, suggesting diminishing returns from the collection of the additional 220+ sources.[33] In contexts where the collection of these additional sources of data is more costly or onerous, there is a value in constraining the number of sources and then employing a statistical correction such as ours. Additionally, the reporting sources naturally classify into three types (i.e., international, national, and regional), with

[31] A greater number of sources may further allow for researchers to introduce explicit correlation parameters for a subset of the reporting sources.

[32] To avoid repetitive citations we note here at the outset that all descriptive statistics presented regarding the ACLED data come from the "ACLED Data Sources," a 2012 working paper linked through in the ACLED event codebook which describes the generation of the data. No individual authors are noted in the text which is available at `http://www.acleddata.com/wp-content/uploads/2014/12/ACLED_Sources-Working-Paper_July-2012_updated.pdf`

[33] Even with many reporting sources, underreporting in social events is still likely to persist. ACLED finds a strong positive correlation between the number of sources and the number of events, even as we increase from 100 to 200 sources, suggesting even with an exhaustive set of reporting sources some underreporting is likely to remain.

the total number of sources in each African country varying from 8 to 100. We believe it is reasonable to think of each reporting type as a 'macro-source' with common misclassification probabilities. For example, as in our paper, international sources are likely to be affected by the amount of reporting coverage generally on a state, the distance to a bureau office, etc. Whereas whether any regional sources reported a true event is more likely to be a function of the number of such sources available to ACLED for that state, which vary considerably, and the degree of press freedom in that state. Considered in this way, we have 3 observed 'reports' – $Y_{INT} = I(Y_{int1} + \ldots + Y_{intK} \geq 1)$, $Y_{NAT} = I(Y_{nat1} + \ldots + Y_{natK} \geq 1)$, and $Y_{REG} = I(Y_{reg1} + \ldots + Y_{regK} \geq 1)$ – each with respective misclassification probabilities – $\alpha_{INT} = \Phi(X,\ \textit{Dist, Coverage})$, $\alpha_{NAT} = \Phi(X,\ \textit{No. of National Sources, Press Freedom})$, $\alpha_{REG} = \Phi(X,\ \textit{No. of Regional Sources, Press Freedom})$. Estimation would then proceed as given in the 3-source variation of the likelihood detailed above.[34]

Beyond the issues on the number of reporting sources, the questions asked by researchers using media-reported event data may differ from what we have introduced here. First, researchers may be interested in event counts (e.g., the number of protests, terrorist attacks, human rights violations, etc.), rather than the binary outcomes we consider here. The derivation of our method above does not apply to these applications, however, our basic likelihood *framework* could be used to address such problems. That is, given a parametric model for the true event-generating process– which we often assume in political science – and a parametric model for the misclassification, we can form the likelihood function of the observed data. Let $\boldsymbol{Y}$, $\boldsymbol{Y}_T$ and $\boldsymbol{X}$ be the reported count, true count, and a set of covariates, respectively. Let $p(y_{\mathrm{T}}|\boldsymbol{X},\beta) = \mathrm{pr}(\boldsymbol{Y}_{\mathrm{T}} = y_{\mathrm{T}}|\boldsymbol{X},\beta)$ be the true data-generating model and $m(y|y_{\mathrm{T}},\boldsymbol{X}) = \mathrm{pr}(\boldsymbol{Y} = y|\boldsymbol{Y} = y_{\mathrm{T}},\boldsymbol{X},\gamma)$ be the model for misclassification. Then one can form an induced model for the reported response $\boldsymbol{Y}$ given $\boldsymbol{X}$, $\mathrm{pr}(\boldsymbol{Y} = y|\boldsymbol{X},\beta,\gamma) = \sum_{y_{\mathrm{T}}} m(y|y_{\mathrm{T}},\boldsymbol{X},\gamma)p(y_{\mathrm{T}}|\boldsymbol{X},\beta)$, write the likelihood, and proceed for estimation of β, the main model parameters.[35]

[34] There is no information loss from our aggregation into clusters over conventional approaches given that the convention in the ACLED data is to aggregate *all* reports into a single binary outcome.

[35] Note that to estimate β, we need γ parameters to be known, and the latter can be estimated from validation data where both the true responses and the reported responses are available for a smaller subset of the original data.

Second, researchers may be interested in using media-based data when determining the sample itself (e.g., studies on the duration of conflicts, whether protests turn violent, etc.). In these analyses, underreporting would result in sample-selection bias – a failure to include a set of true observations in the sample – rather than, or in addition to, misclassification in the outcome. As discussed and presented here our method does not readily address this issue, it is better suited for analyses with fixed observational units. However, despite the well-known methods for handling selection, both generally and with binary-outcome data (Maddala, 1983; Heckman, 1979), none deals with the multiple source issue we have discussed here. As such, we are currently working on a multi-source selection model which would allow for researchers to address these problems.

7 Conclusion

Traditionally researchers devote less attention to measurement error in the outcome, however, here we have highlighted the severity of the bias induced by differential misclassification in binary outcomes. Our simulations show that misclassification can produce substantial bias when researchers employ either: i) strategies which assume no misclassification *or* ii) strategies which assume non-differential misclassification. Further, we show that unbiased estimates can only be obtained by directly estimating a model of misclassification and weighting the risk-model probabilities accordingly.

The threat of systematic measurement error from underreporting is widely discussed in applied research using media-generated event data, yet little work has proposed general strategies to remedy this potential bias.[36] We show how researchers possessing more than one source of data-generating information can achieve this desired result. Specifically, we derive an estimator for applications in which researchers have at least two sources of potentially misclassified data on a single outcome of interest. Under few assumptions, our estimator returns unbiased estimates of the risk probability and allows for source-specific misclassification estimates.

Specifically, we have focused on how our strategy can aid researchers using event-based

[36] A notable exception is Hug (2009).

data comprised from multiple reporting outlets. To our knowledge, no current estimator – in political science, sociology, economics, or statistics – accommodates both multiple sources of reporting data and potential misclassification.[37] Given that many of the first-wave of recommendations to ameliorate reporting bias consisted of gathering data from additional sources, our estimators should reflect this feature of the data-generating process. Yet, as we note, even additional sources are unlikely to result in an uncontaminated data set, meaning that further statistical corrections for misclassificaiton will often be required. As such, we provide a unified method suited for multiple sources of potentially misclassified data. The results show the fitness of our estimator under either differential or non-differential misclassification, suggesting it could be preferred as a general method when researchers are unaware of the nature of the misclassification in their data.

We illustrated the utility of this method in a model of state repression in Africa, observing that predictor effects change dramatically when misclassification is ignored. We believe that similar results will be obtained when researchers utilize our method in studies of protest behavior, civil war, political violence, etc. In future research, we plan to extend on the method introduced here in two ways. First, deriving a semiparametric efficient estimator for the class of problems outlined above. Second, consider cases where the sample itself is defined by potentially misclassified event-based data (e.g., protests, politically-excluded ethnic groups, etc.).

Funding

This work was supported by a grant from the National Cancer Institute [U01-CA057030 to R.J.C.]; and a post-doctoral fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico [201192/2015-2 to B.B.].

[37] The most similar strategy to ours is found in the occupancy modeling literature in ecology, where zero-inflated binomial mixture models are used to estimate detection and occupancy are jointly in biological survey studies (MacKenzie et al., 2006).

References

Abrevaya, Jason and Jerry A Hausman. 1999. "Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells." *Annales d'Economie et de Statistique* 55/56:243–275.

Achen, Christopher H. 2002. "Toward a new political methodology: Microfoundations and ART." *Annual Review of Political Science* 5(1):423–450.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski and Ciprian M Crainiceanu. 2006. *Measurement error in nonlinear models: a modern perspective.* CRC Press.

Carroll, Raymond J and Shane Pederson. 1993. "On robustness in the logistic regression model." *Journal of the Royal Statistical Society, Series B* 55(3):693–706.

Cook, Scott, Betsabe Blas, Raymond Carroll and Samiran Sinha. 2016. "Replication Data for: Two Wrongs Make a Right." doi: 10.7910/DVN/92GMLB, Harvard Dataverse.

Copas, J. B. 1988. "Binary regression models for contaminated data." *Journal of the Royal Statistical Society, Series B* 50(2):225–265.

Davenport, Christian. 2007. "State repression and political order." *Annual Review of Political Science* 10:1–23.

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.

Earl, Jennifer, Andrew Martin, John D McCarthy and Sarah A Soule. 2004. "The use of newspaper data in the study of collective action." *Annual Review of Sociology* 30:65–80.

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.

Heckman, James J. 1979. "Sample selection as specification error)." *Econometrica* 47(1):153–161.

Hendrix, Cullen S and Idean Salehyan. 2015. "No News Is Good News: Mark and Recapture for Event Data When Reporting Probabilities Are Less Than One." *International Interactions* 41(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* Forthcoming.

Hug, Simon. 2003. "Selection bias in comparative research: The case of incomplete data sets." *Political Analysis* 11(3):255–274.

Hug, Simon. 2009. "The effect of misclassifications in probit models: Monte Carlo simulations and applications." *Political Analysis* 18(1):78–102.

Hug, Simon and Dominique Wisler. 1998. "Correcting for selection bias in social movement research." *Mobilization: An International Quarterly* 3(2):141–161.

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science* 54(2):543–560.

MacKenzie, Darryl I, James D Nichols, J. Andrew Royle, Kenneth H Pollock, Larissa L Bailey and James E Hines. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier Academic Press.

Maddala, Gangadharrao S. 1983. *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.

Poe, Steven C and C Neal Tate. 1994. "Repression of human rights to personal integrity in the 1980s: A global analysis." *American Political Science Review* 88(4):853–872.

Poe, Steven C, C Neal Tate and Linda Camp Keith. 1999. "Repression of the human right to personal integrity revisited: A global cross-national study covering the years 1976–1993." *International Studies Quarterly* 43(2):291–313.

Poe, Steven C, Nicolas Rost and Sabine C Carey. 2006. "Assessing Risk and Opportunity in Conflict Studies: A Human Rights Analysis." *Journal of Conflict Resolution* 50(4):484–507.

Salehyan, Idean, Cullen S Hendrix, 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.

Schrodt, Philip A. 2012. "Precedents, progress, and prospects in political event data." *International Interactions* 38(4):546–569.

Schrodt, Philip A and Deborah J Gerner. 1994. "Validity assessment of a machine-coded event data set for the Middle East, 1982-92." *American Journal of Political Science* 38(3):825–854.

Strange, Austin M, Bradley Park, Michael J Tierney, Andreas Fuchs, Axel Dreher and Vijaya Ramachandran. 2013. "China's development finance to Africa: A media-based approach to data collection." *Center for Global Development Working Paper* .

Trumbore, Peter F and Byungwon Woo. 2014. "Smugglers Blues: Examining Why Countries Become Narcotics Transit States Using the New International Narcotics Production and Transit (INAPT) Data Set." *International Interactions* 40(5):763–787.

Weidmann, Nils B. 2014. "On the accuracy of media-based conflict event data." *Journal of Conflict Resolution* 59(6):1129–1149.

Weidmann, Nils B. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1):206–218.

Woolley, John T. 2000. "Using media-based data in studies of politics." *American Journal of Political Science* 44(1):156–173.