Introduction

In 1972, the U.S. and Soviet Union signed the Incidents at Sea (INCSEA) Agreement, which committed both parties to avoid potentially dangerous military interactions between their naval and air forces. Although INCSEA was considered a “quiet success” for arms control during the last two decades of the Cold War, the principal historical analysis of the agreement showed that the reported rate of incidents (the behavior proscribed by the agreement) increased by almost 40% after the agreement was signed.¹ Was the confident initial assessment of the INCSEA agreements effectiveness misplaced? Did the agreement really work to restrain dangerous military activity?

The apparent contradiction between the qualitative and quantitative empirical assessments of the INCSEA Agreement highlights an under-analyzed

aspect of a familiar question: how can we know if an international legal regime is successful in changing state behavior? A large body of the last two decades’ international law and international relations literature indicates that this is a challenging empirical problem. The most prominent difficulty in assessing the effectiveness of international law is endogeneity, or selection effects: the idea that international law, because it ultimately relies on state consent, is epiphenomenal. While endogeneity clearly is a principal obstacle to inference about the effectiveness of international law, another important obstacle – reporting bias – has received much less attention in spite of posing major challenges to the evaluation of effectiveness and compliance.

Reporting bias exists when there is a systematic difference between actual behavior and the recorded data describing that behavior. Many events regulated by international law (such as human rights abuses, trade agreement violations, or border security incidents) may not be perfectly reported – and may have systematically different probabilities of being reported. This difference in reporting probability can be caused by some factor related to the event being reported, the actor responsible for reporting, or the legal agreement itself. Determining the magnitude and direction of reporting bias is often methodologically difficult. Reporting bias can have strong consequences: it can make an effective agreement appear ineffective, or vice versa. Many existing empirical studies of international law explicitly or implicitly assume that compliance and effectiveness data is unbiased without adequate justification.

This paper offers an overview of reporting bias in the context of international law, examines its causes and effects, and identifies some solutions to
improve empirical assessment of international law effectiveness or compliance. The first section advances a conceptual definition of reporting bias along with a number of interdisciplinary examples. The second section explores possible causes of reporting bias, as well as demonstrating that these causes are generalizable to other international legal concepts such as compliance. The third section evaluates the often complex and potentially interactive effects of reporting bias in measuring international law effectiveness. Section four offers a number of strategies for detecting or mitigating reporting bias. Section five provides a more in-depth look at the Incidents at Sea Agreement, examining the causes for reporting bias, application of some techniques for identifying reporting bias and a general assessment of the agreement’s effectiveness in the presence of reporting bias. The sixth section concludes.

1 Conceptually Defining Reporting Bias

The idea behind reporting bias is simple: *some outcomes will be reported less often than other outcomes*. This variability in data collection effectiveness is one of several possible problems with a measurement system, and will introduce a systematic error, or bias, into the resulting data and analysis.

Three factors are required for reporting bias to cause inferential problems:

1. Only a subset of the total population of interest is observed.
2. We do not know the extent of missing observations.
3. Missing observations systematically differ from non-missing observations.

Each of these characteristics is necessary to create the reporting bias problem. Depending on how the observed event data is used, systematic under-
reporting will cause under- or over-estimation of our variables of interest. An example from epidemiology provides a clear illustration of these characteristics and their effects.

Consider the problem of attempting to estimate the incidence or transmission dynamics of a new disease, such as the 2012 outbreak of the previously-unknown Middle East Respiratory Syndrome (MERS).\(^2\) Public health officials must estimate how many new cases of the disease occur (incidence), its contagiousness (transmission), and lethality (how many of those with the disease die from it). In most circumstances, the only data available is reports of disease diagnoses. However, this initial reliance on clinical diagnosis data is known to be problematic, because not all cases of the disease will be reported. Furthermore, the probability of diagnosis (and thus of reporting) is dependent on the severity of the illness. Initially, roughly 30% of those diagnosed with MERS died. One U.S. man who was tested positive for the MERS virus after casual contact with another, more severe MERS case experienced only mild cold or allergy symptoms, which raised the possibility of reporting bias:

The man’s experience suggested that more people may be infected globally with MERS than has been reported, but that their infections went unnoticed because they didn’t get sick, or they developed only mild symptoms. That could mean the MERS virus kills fewer infected people than currently thought. Other emerging infectious diseases like the H1N1 flu also appeared at first to be deadlier than they in fact were, because initially only the sickest cases were diagnosed.\(^3\)

These epidemiological conclusions are still tentative for MERS, but the pro-

\(^2\)McKay 2014.
\(^3\)McKay 2014.
posed effect is a very clear presentation of the dynamic of reporting bias, and the role of the three factors that comprise it.

- **INCOMPLETE OBSERVATION.** The population of interest for public health officials confronting MERS was the total number of people infected with the virus. Not all individuals infected with MERS sought medical attention,\textsuperscript{4} and not all of the MERS cases who did seek care were correctly diagnosed and reported. Therefore, the observed cases were an incompletely observed subset of the full population of MERS infections.\textsuperscript{5} One way of visualizing this is to imagine a spreadsheet of the data: each observation is a row in the spreadsheet table. The observed MERS spreadsheet is missing rows that are present in the full (but unobserved) population spreadsheet of MERS infections.

- **UNKNOWN EXTENT OF MISSING OBSERVATIONS.** In addition to missing observations from our population of interest, we are ignorant of the full size of the population and thus of the scope of our incomplete observation problem. In the MERS example, estimating the reporting rate was impossible because researchers knew the numerator (the number of reported cases) but not the denominator (the total number of infections). To paraphrase Donald Rumsfeld, the extent of missing observations is a known unknown – we usually know or can reasonably suspect that observations are missing, but are unsure of just how many

\textsuperscript{4}Epidemiology terms these “subclinical cases.”

\textsuperscript{5}If observation is complete (all units in the population (or sample frame, in the survey context) are observed but individual variables or characteristics of some units are unobserved), then the situation is one of missing data, which is a different problem from reporting bias. Missing data techniques are discussed in Little and Rubin 2002.
are missing.\textsuperscript{6} Extending the spreadsheet metaphor, we know that rows are missing, but not which ones or the spreadsheet’s total size. Reporting bias thus causes researchers not to know either the proportion of missing observations or which observations are missing. Both are a problem for inference.

- **MISSING OBSERVATIONS DIFFER SYSTEMATICALLY FROM NON-MISSING OBSERVATIONS.** In epidemiology, as in many other fields, missing observations often are missing for a reason – some characteristic of the missing unit causes it to be unobserved.\textsuperscript{7} The U.S. patient with minor MERS symptoms was deliberately tested by public health officials after he was known to have had contact with the more severe U.S. MERS case. In the absence of the active search, the mild-symptom MERS patient would have been an unidentified missing observation. Observed MERS cases are thus probably systematically more severe than unobserved cases, because people suffering mild MERS infections are less likely to seek medical care. Severity of symptoms and access to health care are typical reasons for variation in the probability of observation in epidemiology, but they are often also explanatory variables of interest as well. This correlation between explanatory variables and observability of the dependent variable can have strong and systematic effects on our estimates of the effect of the independent variable on the

\textsuperscript{6}This differs from the problem of survey nonresponse, in which survey researchers, working from a defined sample frame of required observations, are unable to obtain data from some of the designated survey respondents. Survey researchers know both the non-response rate and which observations are missing.

\textsuperscript{7}This effect can be deterministic or probabilistic.
dependent variable.\textsuperscript{8}

MERS is a classic example of the cause and effects of reporting bias. Under-reporting of mild cases of MERS infection in combination with high reporting rates for severe MERS cases resulted in initial underestimates of the number infected with the virus (incidence), underestimates of its transmissibility, and overestimates of its mortality. The remainder of this section offers examples from several other disciplines – including some with closer parallels to international law – to illustrate important aspects of reporting bias.

Measuring crime is another example of potential reporting bias. The primary source of crime data in the U.S. is the Uniform Crime Reports (UCR) system: an FBI compilation of voluntary reports of crimes submitted by law enforcement agencies nationwide.\textsuperscript{9} If we are attempting to evaluate crime trends using UCR data, all three elements of reporting bias are present. Some crimes will be unobserved by police or not reported to them; others will not be reported by police agencies to the FBI. All such crimes are missing observations in the population under analysis. In addition, because the total number of crimes is unknown, the proportion of missing observations is unknown. Finally, there is good reason to suspect that the extent of non-observation varies between jurisdictions, over time, and between types

\textsuperscript{8}If we can establish that there is no systematic difference between missing observations and non-missing observations, the impact of reporting bias can be reduced. The fact that some observations are missing makes this a strong and difficult-to-prove assumption: we lack information on the unobserved group needed to compare it to the observed group. In practice, this usually puts us back at square one, forced to assume that the missing observations are systematically different unless we have a credible reason to believe that the missing-data process is independent of other characteristics.

\textsuperscript{9}Maxfield and Babbie 1998, Chapter 6 and Levitt 1998 discuss crime reporting.
of crimes: missing crimes are thus systematically different from non-missing crimes.

Aviation safety reporting provides another example. The U.S. government operates an Aviation Safety Reporting System to obtain voluntary and confidential reports of air safety incidents\textsuperscript{10} from air traffic controllers and flight crews.\textsuperscript{11} Controllers and flight crews are in good positions to observe incidents, but can face incentives not to report when the incident may reflect unfavorably upon them or their organization. In addition, not all incidents are recognized as such by those involved. Aviation safety reporting thus exhibits the same problems of reporting bias as previous examples from other domains: incomplete observation, an unknown extent of missing observations, and probable systematic differences between observed and unobserved events.

Finally, we consider the problem of measuring hidden populations: a group of individuals for whom the size and boundaries are unknown, and for whom no sampling frame exists ... often involv[ing] stigmatized or illegal behaviour, leading individuals to refuse to cooperate [in providing data] or give unreliable answers to protect their privacy.\textsuperscript{12}

Examples of hidden populations from a study evaluating human trafficking include “prostitutes, traffickers, victims/survivors, or illegal immigrants.”\textsuperscript{13} Empirical analysis of the scope and trends of the human trafficking phenomenon suffers from all three elements of reporting bias discussed above.

\textsuperscript{10}Incidents are defined as events which might have resulted in an accident, but under the existing conditions did not result in damage or injury. Tamuz 1987, p. 72.
\textsuperscript{12}Tyldum and Brunovskis 2005, p. 18.
\textsuperscript{13}Tyldum and Brunovskis 2005, p. 18.
Human rights reporting represents an international law analog to hidden populations: rights abusers and victims both may prefer to conceal their status, while establishing a valid sample frame of rights violations is challenging.

Study of hidden populations almost perfectly recaps the conceptual definition of reporting bias. Observations will almost certainly miss some individuals within the population (imperfect observation). We are ignorant of the true size of the population, so we cannot quantify what proportion of the population has been missed (unknown extent of missing observations). Finally, specific relevant characteristics of interest of the individuals – such as age or immigration status – are likely correlated to their probability of being observed (missing observations differ systematically from non-missing observations.)

This section has defined the concept of reporting bias and identified the three factors that make reporting bias an inferential problem: incomplete observation, unknown extent of missing observations, and a systematic difference between missing and non-missing observations. We now turn to examining potential causes of reporting bias.

2 Causes of Reporting Bias

The causes of reporting bias can be very complex and context-specific. This paper uses a generic three-stage model to describe potential causes for reporting bias. For purposes of illustration, the model assumes a process of reporting violations of some international obligation (such as mandated economic sanctions). Three participants are involved: an actor, who can engage
in proscribed behavior; an observer (or agent) who is assigned to detect and report violations by the actor; and the state (or principal), who supervises and receives reports from the agent, and provides external reports of violations to other states or an international organization. Discussion of each stage illustrates typical sources of bias in reporting.

STAGE 1: ACTOR/EVENT. In the first stage, the actor causes events that violate international law (such as trading with a sanctioned state).\(^\text{14}\) The event-generating process can be random (unaffected by the possibility of detection), strategic (affected by the possibility of detection and reporting), or a mix of both random and strategic processes. One possible source of reporting bias occurs at the event stage: will the event be detected by the observer? Events that are not detected cannot be reported. The *inherent observability of the event of interest* is the first major cause of reporting bias. Observability can vary: small-scale or less severe events are often harder to detect than large-scale or severe events. Actors may have incentives to conceal the events from observation, particularly if there is stigma or illegality that can negatively impact them. The probability of observation can affect strategic event behavior by the actor, but will not impact random event behavior.\(^\text{15}\) Non-detection of relevant events induces reporting bias: in the example context of event reports of violations, the violation counts and rates will be biased low, while a compliance standard rating will be biased high (in the direction of better compliance than is actually appropriate).

\(^\text{14}\)The actor need not be an actual person – it could be a random process that causes a state to fail to meet its international legal obligations.

\(^\text{15}\)Note that a change in event probability caused by the actor’s strategic response to the possibility of detection also changes the ultimate reporting output, but this is the hypothesized effect of interest: the international agreement is effective in changing behavior.
STAGE 2: OBSERVER/AGENT. In the second stage, an observer aware of an event must evaluate if the event meets reporting criteria, decide whether or not to report it, and then successfully submit a report. Possible causes of reporting bias at this stage include the following: agent incentives to report, under-report, or over-report; the clarity of reporting criteria (exactly what events justify or require a report); the effectiveness of reporting channels, and agent perceptions of the principal’s ability to detect an inaccurate report. Biased internal reporting by agents can thus have deliberate (suppression or inflation of event counts due to agent-level incentives) or inadvertent causes (confusing or vague reporting criteria, noisy or ineffective communications channels, competent or incompetent agents). Internal reports of event counts, rates, or standards thus may not accurately reflect agent observations.

STAGE 3: STATE/PRINCIPAL. In the third stage, the state (or principal)\textsuperscript{16} receives reports submitted by agents, evaluates them, and must decide what external reports to submit. The causes of reporting bias at the state level are essentially similar to the observer/agent level: reporting incentives, ability to detect inaccurate reporting by the agent, clarity of criteria, and reporting channel effectiveness.\textsuperscript{17}

This three-stage actor/agent/principal reporting model captures a number of potential causes of biased reporting. These can usefully be summarized in the following categories:

- Observability: can the event of interest (or the completeness and accu-

\textsuperscript{16}The state represents the applicable government agency or official responsible for collecting and forwarding reports of violations to other parties to the agreement or a cognizant international organization. Reported data ultimately comes from the principal.

\textsuperscript{17}Reporting channel effectiveness can represent both technical efficiency of communication and organizational capacity – bureaucratic process efficiency, corruption, etc.
racy of reports thereon) reliably be detected by the agent or principal?

- Clarity: are evaluation criteria clear and unambiguous as to whether or not specific behavior qualifies as legally compliant or non-compliant?

- Incentives: do actors, agents, or principals have incentives that would encourage them to conceal events, or to under- or over-report them?

- Capacity: are reporting processes technically effective, and do agents and principals have the expertise and resources to make them work?

These causes are all logically independent. Any single cause can result in biased reporting; multiple causes also can potentially reinforce or cancel out each other. Additionally, the working of a cause at one level can have an interactive effect at another level of the model. For example, an agent that knows its principal has good direct observability of events may feel obliged to report it accurately even if it fears that a report of the event will reflect poorly on it.\(^{18}\) Conversely, an actor who is aware that the relevant agents and principal lack the capacity to detect and report violations may feel emboldened to commit violations.

The basic model can be adapted easily to other detection and reporting contexts: multiple parties (such as two rival states monitoring an arms control agreement, or a multilateral trade agreement) or different numbers of levels (a longer or shorter reporting chain). The basic potential causes of reporting bias – observability, clarity, incentives, and capacity – remain applicable regardless of the specific reporting context.

\(^{18}\) Tamuz 1987, pp. 70-71 discusses this dynamic in the context of aviation safety reporting.
3 Effects of Reporting Bias

When reporting bias is present, certain outcomes of interest will be reported less frequently than other outcomes. There is systematic error in the outcome data – some portion of the event distribution will be under-reported.\textsuperscript{19} This distortion of the underlying distribution of outcomes affects the measurement and assessment of the effectiveness of international law in complex and potentially interactive ways.

The effect of biased reporting manifests itself in measured variable output. Observation and reporting of behavior is the first step in assessing; measurement is the next. Measurement takes a reported observation and assigns some value (generally numeric) corresponding to the observed behavior to a variable. The measurement depends on both the variable type being used as well as the nature of the behavior being observed. The assigned value may differ from the underlying behavior.

The type of variable coding used affects the impact of reporting bias. Variables are structured on a continuum of increasing informational content: from binary (yes/no), through ordinal (ordered scales), to interval (quantitative data).\textsuperscript{20}

Logically, it doesn’t matter if we are looking at compliance (performance of obligated behavior or non-performance of prohibited behavior) or at effectiveness (behavioral or other evidence of achieving an obligated outcome).

\textsuperscript{19}Over-reporting will have the opposite effect, but the conceptual application is similar. In addition, it should generally be easier to identify and rule out bogus additional reports than to identify required reports that are missing.

\textsuperscript{20}This discussion of variable types omits the nominal (unordered categorical) type. The binary type is often considered as a two-category nominal variable.
For both, we are ultimately counting something: events that happen (such as the number of business transactions in violation of a sanctions regime, or unlawful attacks on civilians) or characteristics of events that happen (the dollar value of a sanctions-violating transaction, or the number of civilians killed in an unlawful attack). Sometimes these counts will be used either directly or in aggregation (the total number of sanctions-violating transactions; the total value of transactions) but often they will be transformed into binary, index, or ordinal variables. Finally, continuous interval variables such as rates, percentages or proportions can be used. The effect of reporting bias can vary depending on the extent of missing data and on how the data is applied in the measurement process.

The next section turns to addressing each type of variable and how reporting bias is likely to affect it.

Count variables

Count variables are conceptually simple: they record the number of events meeting certain criteria within a specified period, or the duration of an event or process, measured discretely. Typical examples in studies of international agreements include:

- The total number of noncombatants intentionally killed by each side during a war.\(^{21}\)
- The number of states ratifying a particular treaty, either globally or within a particular region.\(^{22}\)

\(^{21}\)Valentino, Huth and Croco 2006, p. 359.
• The number of years since the last current account restriction in violation of Article VIII of the International Monetary Fund Treaty prohibiting such restriction.\textsuperscript{23}

• The number of maritime incidents involving U.S. and Soviet naval forces in a given year.\textsuperscript{24}

For counts, reporting bias (under-reporting of some or all of the events) will result in under-counting. The reported count will be systematically biased downwards. Consider the example of intentional civilian battle deaths. The intentional death count $D$ is the sum of civilian deaths $d$ in each of a total of $n$ engagements:

$$D = \sum_{i=1}^{n} d_i$$

where $d$ is the number of deaths in each engagement.

This measure is almost certainly biased downwards: we don’t have the death count $d_i$ for all engagements; and we frequently may not know the total number of engagements $n$. As a result, the reported count $D$ of deaths understates the unknown true count. Of course, it is also possible for death counts to be erroneously high, but this would be an example of measurement error, not reporting bias per se.

**Binary variables**

A binary (also called a dichotomous or dummy) variable has two possible values.\textsuperscript{25} The legal concept of compliance naturally codes as a binary variable.

\textsuperscript{23}Von Stein 2005; Simmons 2000.
\textsuperscript{24}Ling 2012; discussed further below.
\textsuperscript{25}These can be coded as Yes or No, True or False, or as 1 or 0, but all forms are logically equivalent.
– a state’s actions are consistent with its international legal obligations, or they are not. Binary variables are frequently used in empirical studies of international law. Representative examples include:

- Ratification of a human rights treaty used as a proxy for formal acceptance of international human rights law.\(^\text{26}\)

- Compliance or non-compliance with a third-party arbitration or adjudication decision (non-compliance defined as a decision that was rejected, protested, or ignored by one party to the dispute).\(^\text{27}\)

- Restriction on payments in current account (compliance with Article VIII of the International Monetary Fund Treaty prohibiting such restrictions of currency exchange in current account.).\(^\text{28}\)

In practical measurement, a single instance of non-compliant behavior can be enough to determine non-compliance with the obligation. Consider binary measurement of a state’s compliance with its obligations under the Convention Against Torture. The binary compliance variable \(B\) would receive a value of 0 for a single documented instance (or more) of torture. \(B\) would receive the value 1 (representing compliance) only if the count \(t\) of torture events was 0:

\[
B = \begin{cases} 
0 & \text{if } t \geq 1 \text{ where } t \text{ is the count of torture events} \\
1 & \text{if } t = 0 
\end{cases}
\]

Binary variables do not distinguish between a single instance of non-performance and a massive pattern of violations: both are coded identically as non-

\(^{27}\)Simmons 2002, p. 850.
\(^{28}\)Von Stein 2005; Simmons 2000.
compliance with the international legal obligation.\footnote{Note that a binary variable uses a threshold value to classify the result – the threshold need not be set at 1 as in this example.}

Binary coding of compliance or effectiveness is more resistant to reporting bias than a count variable. In cases where the underlying count is relatively large, it should be unlikely that reporting bias will drive the reported count to zero and force a change in the binary output variable from one to zero. However, if the underlying count is relatively small and under-reporting could plausibly drive the reported count to zero, a binary variable can still suffer from reporting bias.\footnote{The length of periods under analysis (events per month or per year) can affect a binary variable’s sensitivity to reporting bias. A longer period (because of a longer expected count) would probably be less susceptible to reporting bias because its underlying count would be less likely to reach zero.}

Another implication of the binary variable structure is that a single undetected false positive report\footnote{In this example, a single instance of a false claim of torture.} will cause the resulting binary output to be biased in the direction of non-compliance.

**Ordinal variables**

Ordinal variables capture an ordered categorization. Many variables of interest to scholars of international law are ordinal, with multiple categories that have a logical ordering or ranking. The following examples illustrate different applications of ordinal variable coding:

- Freedom House’s Freedom Status variable: Free, Partly Free, or Not Free.\footnote{Freedom House 2014.}

- The Cingranelli-Richards (CIRI) Human Rights Data Project’s Political or Extrajudicial Killings variable: 0-practiced frequently, 1-practiced occasionally, or 2-have not occurred/unreported.\footnote{Cingranelli, Richards and Clay 2014.}
The five-level Political Terror Scale: ranging from 1 (secure rule of law, no political imprisonment, rare torture, extremely rare political murders) through 2 to 3 (extensive political imprisonment, common political murders) and through 4 to 5 (level 3 practices expanded to the whole population).\textsuperscript{34}

The Polity IV Dataset’s Polity Fragmentation variable: 0-no overt fragmentation, 1-slight fragmentation, 2-moderate fragmentation, 3-serious fragmentation.\textsuperscript{35}

Many ordinal variables (such as the Polity Fragmentation variable) reflect an underlying count or interval variable (e.g. percentage of national territory actively separated from central regime authority) that do not have equal unit intervals. For the Polity Fragmentation example, the category 1, 2, and 3 values cover 10%, 15%, and 25%-wide ranges of territorial non-control respectively.\textsuperscript{36} Thus, the difference between Polity Fragmentation values of 0 and 2 could be substantially smaller than the difference between values of 1 and 3.

Other ordinal variables (such as the Political Terror Scale) lack an underlying interval value and are constructed more qualitatively. As with the interval-derived ordinal variables, the differences between levels can vary from one pairing to the next.

The effect of reporting bias on ordinal variables is harder to predict. A small to moderate bias in an underlying count or interval variable can cause

\textsuperscript{34}Poe, Carey and Vazquez 2001, p. 658; Wood and Gibney 2010.
\textsuperscript{36}The 1 value (slight fragmentation) indicates that fragmentation is greater than zero but less than 10%; a 2 (moderate) fragmentation score reflects 10-25% of territory not under central authority, a 3 (serious) fragmentation score reflecting 25-50% of territory not under central authority. There is an additional implied higher category of fragmentation (> 50% of territory not under central control) that is captured in a separate binary State Failure variable.
a shift from one ordinal value to the next in some cases, but not in others. The effect of a given amount of reporting bias may depend on where (relative to the category cutpoints) a particular observation lies.\textsuperscript{37}

**Rate and proportion variables**

Rates and proportions are a naturally bounded set of values: typically, they range from 0 to 1 (for proportions) or 100 (for percentages). Rates and proportions are defined as the number of events meeting specific criteria divided by the total number of relevant events.

Examples of rates and proportions in international law compliance scholarship include:

- The measurement of women’s political equality by the percentage of men in a state’s legislature.\textsuperscript{38}
- The measurement of the strength of a regional legal norm by the proportion of regional states who have accepted a specific international legal obligation.\textsuperscript{39}

Consider a variable measuring the proportion of noncombatants intentionally killed in relation to the total number of intentional combatant and noncombatant battle deaths:

\[
P = \sum_{i=1}^{n} \frac{nc_i}{nc_i + c_i}
\]

\textsuperscript{37}For example, the same 5\% understatement of the amount of territory not controlled by the central government for the Polity Fragmentation variable may be much more likely to cause a category change in the narrower categories 1 or 2 than in the wider category 3.\textsuperscript{38}Hathaway 2002, pp. 1975-76.
\textsuperscript{39}Simmons 2000, Von Stein 2005. Simmons’ Regional Noncompliance variable similarly measures the proportion of IMF members within a region who restrict their current account in violation of an Article VIII legal commitment not to do so.
where \( nc \) is the number of noncombatant deaths and \( c \) is the number of combatant deaths in each engagement \( i \).

Rates and proportions can be affected by reporting bias in a number of ways. In the proportion of intentional noncombatant deaths example, suppose that small engagements are less likely to be reported than large ones. If the ratio of noncombatant to total deaths is constant regardless of engagement size, then the recorded proportion of noncombatant deaths will be unbiased in spite of the fact that we have not recorded all of the relevant events. However, if large engagements kill a higher proportion of noncombatants than smaller engagements do, the biased reporting that fails to capture all smaller engagements will cause the recorded proportion of noncombatant deaths to be biased upwards.\(^{40}\)

**Index variables**

Index variables are a combination of several separate variables, with the values of the individual components summed to form the value of the index variable.

\[
X = \sum_{i=1}^{n} V_i \text{ where } V_i \text{ is one of a series of } n \text{ variables.}
\]

Examples of index variables in international law compliance include:

- Wade M. Cole’s composite index of respect for women’s rights constructed from the CIRI Human Rights Data Project’s separate variables for Women’s Economic Rights, Political Rights, and Social Rights.\(^{41}\)

\(^{40}\)Conversely, if large engagements kill a lower proportion of noncombatants, then our recorded variable will be biased downwards, understating the true proportion of noncombatant deaths.

\(^{41}\)Cole 2013, p. 236. CIRI discontinued the Social Rights variable in 2005.
Each of three components had a 0 to 3 ordinal score; Cole’s composite index variable summed the three component variables, producing a range of ordinal scores from 0 to 9.

- The Freedom House political rights score (the lowest tier in the three-tier Freedom House rating system). A country is awarded 0 to 4 ordinal points for each of 10 political rights indicators. The total political rights score is the sum of the 10 indicators, with scores ranging from 0 to 40. This indexed score is used as the basis for assigning a political rights rating.

Index variables, as the sum of multiple individual variables, may be more resistant to reporting bias than their component parts – unless the reporting bias affects most or all components of the index.

**Evaluating the effects of reporting bias**

To return to the broader concept of the effects of reporting bias, it is necessary to consider both the expected nature of the biased reporting (which types of outcomes are less likely to be reported) and how that biased reporting will impact the variable type being used to measure the outcome. Count variables typically will be biased downwards. Binary variables will be biased less often, unless reporting bias is severe enough to push the reported count below the binary threshold value. Ordinal variables must be examined to evaluate whether or not reporting bias is likely to result in a change from one category to another. Rate and proportion variables must be evaluated to determine how biased reporting will affect the numerator and denominators of the variable; the effect will vary depending on where the under-reporting fits into the fraction. Finally, index variables may be biased downwards, but

---

42Freedom House 2014.
this effect will depend on how strongly reporting bias affects the components of the index.

4 Strategies for Dealing with Reporting Bias

The previous discussions of the concept and causes of reporting bias suggest that reporting bias cannot be eliminated in many contexts relevant to scholars of international law effectiveness or compliance. The challenge for the empirical scholar is threefold: to recognize the presence of reporting bias, to mitigate its effects when possible, and to assess its impact when mitigation is not possible.

Recognition of reporting bias is the first step. Once the variable of interest is defined, the analyst must evaluate the likelihood of relevant but unreported events. In some cases, there will be very high confidence that all relevant observations have been captured in the data. In others, we may suspect that reporting bias exists, but may be uncertain of its extent or potential impact. In yet others, we will be confident that the data suffers from biased reporting.

Mitigating the effects of reporting bias is the next step, but it is not always possible. In some cases, additional research can identify missing reports. This might involve examining the archives of lower-level units of government (the observer/agent of the three-stage model discussed in section two above) or of the main government agency responsible for submitting external reports (the state/principal). It can also involve aggregation of multiple sources of information – comparing the U.S. State Department and Amnesty International human rights reports, for example. In other cases, retrospective mitigation
may not be possible but a recognized reporting bias problem might justify additional resources (to build reporting capacity or to strengthen agent incentives to report) that could prevent bias in future reporting.

Assessing the impact of reporting bias is the final task. If mitigation is not possible, it may be the only option. Assessment of reporting bias essentially involves asking “how would the data look if reporting bias were not present?” It is counterfactual in the sense that the “unbiased” data is conjectural. However, such conjecture need not be entirely detached from reality. There may be sound theoretical or empirical reasons for regularity in the data that can be exploited to arrive at a plausibly rigorous counterfactual value for the variable of interest. Alternative, it may be possible to create a bounded interval estimate of the variable of interest.

Three strategies can help with these challenges: comparing multiple information sources, applying an expected underlying data distribution, and exploiting events that are always reported. While not all of the examples listed below are from international law contexts, they do illustrate how these approaches can work to identify, mitigate, and quantify the scope of reporting bias.

**Comparing multiple information sources**

Multiple independent sources of information can help identify reporting bias in one or both sources used if the various causes of reporting bias differ between them. Sampling methods can be a particularly effective tool in this respect if the nature of the data permits their use.

For example, consider the problem of estimating the U.S. crime rate.
Two independent approaches have been used: collection of police reports (the Uniform Crime Reports (UCR) data discussed above in section 1) and an annual survey of roughly 100,000 individuals in U.S. households (the National Crime Victim Survey, or NCVS). Because the UCR only reports crimes known to the police, the NCVS methodology bypasses the police to ask residents directly about their experiences with crime. The NCVS results reflect a significantly higher rate of crime than the UCR results, although the UCR data is more specific in some respects.43

Another example of multiple source analysis using sampling is drawn from studies of hidden populations, which use a two-stage “capture-recapture” sampling approach. The capture-recapture method involves taking a first-stage sample, and then determining how many first-stage observations are “recaptured” during the second stage. The recapture rate enables an estimate of the probability of detecting the presence of a specific individual with a specific characteristic in a single sample, and of estimating the total size of a population of interest.44

**Exploiting events that are always reported**

While reporting bias is a frequent problem, it is not universal. In particular, certain types of events may be always, or virtually always, reported. Some openly public variables are in the “always reported” category – such as a treaty signature or ratification. Other events that will typically never go unreported include wars, restrictions on the current account, or debt defaults.

Events that are always reported should be immune to reporting bias –

---

44Tyldum and Bruonovskis 2005, p. 29.
a fact that can be exploited to evaluate possible bias in similar events that are not always reported. The economist Steven Levitt developed this approach in analyzing domestic crime data: he assumed that since murders are almost always reported to police, reporting bias in other crimes would make the ratio between murder and other violent crimes vary from jurisdiction to jurisdiction.\textsuperscript{45} Later in this paper, I apply Levitt’s ratio technique to one aspect of U.S.-Soviet naval incident data.

**Apply an expected underlying distribution**

In some cases, the population of reportable events may have an empirical regularity that can be exploited. An expected underlying data distribution can be combined with the available reported data to evaluate the extent of reporting bias. Where a less-biased portion of the underlying distribution is available, the under-represented or missing data can be estimated using the distribution.

An example of this process is the application of a power law distribution in evaluating the frequency and severity of violent events (such as conflict or terrorism deaths).\textsuperscript{46} Power laws describe data where the logarithm of an event’s severity and the logarithm of its probability form a straight line when plotted together. Jeffrey Friedman applied the power law distribution to the best historical record of casualties from individual engagements in the “Indian Wars” between Native Americans and the U.S. Army from 1776 to 1890.\textsuperscript{47} Based on the expected power law distribution, Friedman was able

\textsuperscript{45}Levitt 1998, p. 64.
\textsuperscript{46}This example is drawn from Friedman 2014.
\textsuperscript{47}Friedman found that while the probability and severity of U.S. casualties per engage-
to extrapolate a total of roughly 50,000 Native American and roughly 12,000 U.S. casualties (against recorded totals of roughly 26,000 Native American and 10,000 U.S. casualties).\textsuperscript{48}

The doubling of the Native American casualty estimate shows how information about an expected distribution of data can enable extrapolation of unreported data. While such estimates must be used with care (and very clearly identified as estimates or extrapolations), appropriate application of an expected underlying distribution can provide a powerful tool for identifying and correcting reporting bias.

With a set of strategies for identifying, mitigating, and assessing reporting bias in hand, we turn to applying them in an empirical evaluation of the effectiveness of an important but little-known Cold War international agreement – the 1972 U.S.-Soviet Agreement for the Prevention of Incidents On and Over the High Seas.

5 Reporting Bias in the INCSEA Agreement

The 1972 Incidents at Sea (INCSEA) Agreement between the U.S. and Soviet Union was one of the earliest international agreements that imposed mutual constraints on peacetime operational behavior of military forces. A large but purely qualitative literature has long considered INCSEA to be one of the most successful of the Cold War’s military confidence building measures.

\textsuperscript{48}Friedman 2014, pp. 2-4.
This section examines a quantitative analysis of an original dataset of naval incidents from 1960 to 1988 to evaluate if the agreement’s performance in reducing naval incidents matches the positive qualitative assessment. While the quantitative results are mixed, there is substantial evidence of reporting bias in the data. Application of one reporting bias mitigation approach supports the positive qualitative assessment of the INCSEA Agreement’s effectiveness.

**Incidents at Sea: Problem and Agreement**

The early Cold War saw relatively little direct interaction between the navies of the U.S. and the Soviet Union, primarily because the Soviet Navy was small and remained in home waters. However, during the 1960s the USSR rapidly expanded the size and deployment posture of its fleet. By the end of that decade both superpowers were accusing the other of reckless maneuvering, harassment, or threatening behavior that risked collision or armed escalation. These cases of dangerous behavior became collectively known as “incidents at sea.”

By one count, in 1969 there was on average one collision or near-miss per month between Soviet and American ships and aircraft. While some of these incidents were likely caused by poor judgment or overzealous local commanders, others may have represented deliberate harassment sanctioned by higher military authorities.49

In 1971 and 1972, two series of negotiations reached quick agreement on a variety of measures to prevent such incidents. The Agreement was signed

49McVadon 2009, pp. 7-8.
in Moscow on May 25, 1972.\textsuperscript{50} As signed, the INCSEA Agreement had the following key provisions, all of which remain in effect today.\textsuperscript{51}

- Reaffirmed the obligation of Soviet and U.S. ships to follow the nautical Rules of the Road.
- Added prohibitions against hindering formations of ships, aircraft launch and recovery, and replenishment operations – situations unique to naval vessels that are not specifically addressed in the Rules of the Road.
- Regulated conduct by military aircraft, for which internationally recognized rules did not exist.
- Prohibited specific forms of harassment, including simulated attacks, aiming weapons or launching objects at other ships, shining searchlights at other ships at night, and aerobatic overflights of ships.
- Provided a signaling system for ships and aircraft in close proximity to communicate intentions, as well as requiring advance notification of potentially hazardous exercises such as missile test firings.
- Created an annual review process, where alleged violations and other matters of relevant concern could be discussed by naval representatives of both sides. In addition, the parties agreed that information on incidents would be exchanged through their respective naval attaches rather than diplomatic channels.

The original agreement has survived with little substantive change.

**INCSEA Effectiveness - A Qualitative Take**

The INCSEA agreement was intended to provide a sound basis for continued conduct of the U.S.-Soviet military competition at sea but with procedures to reduce misunderstanding and the most dangerous forms of activity. Most

\textsuperscript{50}Lynn-Jones 1985, pp. 173-75.  
\textsuperscript{51}Lynn-Jones 1985, pp. 174-75.
qualitative assessments agree that the INCSEA agreement was effective in reducing the occurrence and severity of at-sea incidents between the naval and air forces of the two superpowers. Statements by high-level naval officials of the USSR and United States both supported this overall assessment, and particularly emphasized the value of the annual review consultation mechanism. Most scholarly and policy analysts shared the official view.

Sean Lynn-Jones produced the canonical analysis of the INCSEA agreement in 1985. His title neatly summarizes his conclusion: “A Quiet Success for Arms Control.” Almost all post-1985 discussion of the INCSEA agreement draws heavily on Lynn-Jones’ research. Lynn-Jones provided a summary of incident behavior and its incentives, a history of the INCSEA negotiations, and an account of the agreement’s implementation. His assessment was favorable, and is worth quoting at length because it has been so widely cited in the literature:

The [INCSEA] agreement has helped to avert potentially dangerous incidents between the U.S. and Soviet navies. Most of the achievements of détente have lost their luster with the passage of time, but the agreement’s effectiveness appears to have survived the deterioration of U.S.-Soviet relations. . . . Since the agreement was signed, fewer serious naval confrontations have occurred, and those that take place have not generated dangers of escalation or political crises.

Can these positive assessments be supported by quantitative analysis of...
the best available unclassified data? The next section turns to this question.

**INCSEA Effectiveness: A Quantitative Take**

This section evaluates the INCSEA agreement’s effectiveness by comparing the likelihood of incidents before and after the agreement was signed. The dataset used is an original one derived from a published chronology of incidents based on U.S. Navy archives. This section of the paper presents the analytical concept, the key variables, and an evaluation of the results. More detailed information on the data sources, their quality, the statistical methodology used and the detailed statistical results are in the appendix.

**Quantitative Hypotheses**

The INCSEA Agreement was designed to prevent the most dangerous incidents: the outbreak of hostilities at sea, death or injury of military personnel, and damage to or loss of ships and aircraft. The historical record shows that there were confirmed cases of death, loss of aircraft, and damage to ships before the agreement. After the agreement was signed in 1972, there were no hostile actions, no deaths or injuries, and no loss of ships or aircraft directly attributable to interactions between U.S. and Soviet forces.\footnote{Incidents resulting in minor damage to ships or aircraft did still occur after the agreement was signed.}

If the INCSEA agreement was in fact effective at achieving these objectives, at-sea incidents should have been both less frequent and less severe after the agreement was signed than before. This reasoning leads to two hypotheses:
• H1: The probability of an incident escalating to a serious level (involving collision, death, injury, or loss of a ship or aircraft), conditional on an incident occurring, should decrease after the INCSEA Agreement was reached, other things being equal.

• H2: The probability that an incident occurs at all should decrease after the INCSEA Agreement was reached, other things being equal.

Data

The dependent variable is the naval incident rate while the independent variable is whether or not the INCSEA agreement was in effect. There are three control variables used in the event count regression model: the overall quality of U.S.-Soviet relations, the density of at-sea interaction between the two nations’ naval and air forces, and the time elapsed since the agreement was signed.\(^5^7\)

The primary data source for the dependent variable (observed incidents) is a declassified summary of incidents derived from the files of the principal U.S. Navy staff office responsible for INCSEA agreement negotiations and compliance monitoring. They were declassified in the late 1990s and used to create a detailed published chronology of U.S.-Soviet naval incidents, which

\(^{57}\)The incident probability would logically be likely to increase when U.S.-Soviet tensions were high. Increased risk of war in the Cold War naval context provided a number of incentives for more risky or provocative behavior. Similarly, the probability of incident would logically increase with higher densities of U.S.-Soviet naval interactions. More ships and aircraft in proximity would offer greater scope for accidents and misjudgment, as well as possibly increasing mutual suspicion and tension. Finally, the time elapsed since the agreement might influence the incident rate in two ways. In the first case, as both Soviet and U.S. forces gained familiarity with the INCSEA agreement and its provisions, the incident rate might be expected to decline over time. In the second, the agreement’s effectiveness in reducing incidents might have peaked early and then declined over time as attention to the agreement slackened or as accumulations of agreement violations degraded the cooperative equilibrium.
I developed into an original dataset.\textsuperscript{58}

The dataset contains 463 incidents that occurred between January 1960 and December 1988, categorized into two types: maneuvering and non-maneuvering incidents. In addition, serious incidents make up a third, cross-cutting category. The categories are briefly summarized here:

- **Maneuvering Incidents** involve the creation of collision risk or physical harassment by the movements of ships or aircraft in close proximity. There were 323 maneuvering incidents in the dataset.

- **Non-Maneuvering Incidents** create risk of unintended escalation or physical hazards not involving ship or aircraft movements. There were a total of 140 non-maneuvering incidents in the dataset.

- **Serious Incidents** are a cross-cutting classification for any at-sea interaction between the superpowers’ ships and aircraft that resulted in death, collision, or loss of a ship or aircraft. In theory, serious incidents could be either maneuvering or non-maneuvering events, but the nine actual serious incidents were all maneuvering incidents: eight collisions and one Soviet aircraft that crashed while buzzing a U.S. ship.\textsuperscript{59}

The maneuvering/non-maneuvering incident distinction distinguishes between behavior that is dangerous through the risk of collision and other behavior that is primarily dangerous through creating the fear of attack. Maneuvering incidents, because they were at least partially based on a pre-existing foundation of international law, tended to be reported both before

\textsuperscript{58} Winkler 2000, pp. 177-210 is the published source. Winkler’s information was derived from the official records of the Oceans Policy Branch of the Office of the Chief of Naval Operations, currently housed in the U.S. Navy Operational Archives.

\textsuperscript{59} All eight collisions (surface and air) were the maritime equivalents of automotive “fenderbenders,” involving only minor damage. However, warship collisions were potentially extremely dangerous. A 1969 collision between an Australian aircraft carrier and an American destroyer resulted in the sinking of the destroyer, killing 74 of its crew. \textit{New York Times}, June 4, 1969, p. 1. The implications of such an accident involving U.S. and Soviet warships was one of the major drivers of the INCSEA agreement.
and after the INCSEA agreement was signed. Non-maneuvering incidents were much less likely to be reported before the agreement. Finally, serious incidents are based on an objective and easily measured standard for the most dangerous outcomes – one that was virtually certain to be reported.

The incident dataset thus provides a systematic coding of the best available open-source compilation of U.S.-Soviet naval incidents for a 29-year period spanning the most intense naval activity of the Cold War.

Figure 1 illustrates the incident data, showing a time series of incidents. From top to bottom, the plots show all incidents, maneuvering incidents, non-maneuvering incidents, and serious incidents. The vertical dotted line represents May 1972, when the INCSEA Agreement was signed. The data are noisy: over one-third of the 348 months in the dataset had no incidents, while five months had 10 or more incidents each. Summary statistics are included in the data appendix in Tables 1 through 4. I use two incident counts as dependent variables: maneuvering incidents (MANINC) and all incidents (ALLINC).

While comprehensive and based on authoritative source material, this dataset undoubtedly contains some biases. There are definite trends in the data that suggest particular types of behavior either occurred or were reported with different frequencies at different times. Also, there is evidence in the U.S. Navy records that some filtering of reports occurred, probably at multiple levels of the chain of command. If filtering or reporting trends oc-

---

60 The third spike from the left on the serious incident plot is actually two incidents, one in each of the two successive months, December 1967 and January 1968. Also note the different scale for the relatively infrequent serious incidents.

61 For instance, reports of radar illumination and gun pointing at ships and aircraft were particularly prevalent from 1975 to 1985, but infrequent at other times.
Figure 1: Monthly Incident Counts, 1960-1988
curred consistently throughout the evaluation period, then these errors might self-cancel; if they were not consistent, the data will be biased.

One of the inherent challenges of evaluating the impact of international agreements is that they tend to affect both behavior and reporting. The INCSEA Agreement is no exception. While the INCSEA Agreement may have encouraged both sides to report systematically behavior that had previously been reported more irregularly (particularly, maneuvering incidents by ships), it also created new categories of proscribed behavior (such as maneuvering incidents by aircraft, and most non-maneuvering incidents). The creation of new, previously undefined, prohibited behavior is one of the primary reasons for making the distinction between maneuvering and non-maneuvering incidents.

INCSEA Effectiveness: Quantitative Results

The INCSEA quantitative analysis is presented in three parts. First, I evaluate the descriptive statistics of incidents by type before and after the agreement. These provide a rough estimate of the relative frequency of serious and non-serious incidents, and also suggest that either reporting standards or incident behavior (or both) changed after the agreement was implemented. Second, I briefly summarize the results of the regression model testing Hypothesis 2 – which suggest that incident rates increased after the INCSEA Agreement was signed. Finally, I evaluate for the presence of reporting bias and check whether reporting bias can explain the discrepancy between the pessimistic regression results and the positive assessment of the agreement’s effectiveness in the qualitative literature.
Examining the Descriptive Statistics

The descriptive statistics presented in Table 2 raise several important points.

First, serious incidents, although always infrequent, were much more likely to occur before the INCSEA Agreement than after: the monthly serious incident rate (the mean in Table 2) declined almost fivefold. These figures directly support the statements by Soviet and U.S. officials that the most dangerous behaviors were greatly reduced by the agreement.

Second, non-maneuvering incidents were very infrequent before the agreement but much more prevalent afterwards, especially between 1976 and 1983. This is logical because these behaviors were not restricted by any international rules before the INCSEA agreement. Before 1972, these activities might have been considered dangerous or rude, but they did not violate either the letter or spirit of any international obligations. After 1972, they were explicitly covered by the INCSEA agreement. This suggests that the increase in reported non-maneuvering incidents is an example of reporting bias: the agreement defined such behavior as violations of the agreement and therefore prompted more reporting. The alternative explanation – that the non-maneuvering incident behavior increased dramatically after the INCSEA agreement was signed in 1972 – is much less plausible.

Third, the relative paucity of non-maneuvering incidents before 1972 makes the total number of incidents a biased measure of incident behavior. The maneuvering incident record shows much more consistency before

---

62 Non-maneuvering incidents, to recap, included weapon pointing, radar illumination, mock attacks, and dropping objects from aircraft.

63 One partial exception is the pointing of weapons at aircraft – which was not explicitly prohibited by the INCSEA agreement, although both sides agreed that this behavior violated the agreement’s spirit.
and after the agreement. Complaints of dangerous maneuvering behavior\footnote{Maneuvering incidents included buzzing, interference with flight operations or underway replenishment, and failure to comply with the Rules of the Road.} are well represented both before and after 1972. For this reason, I consider that the maneuvering incident count is a less biased dependent variable than the total incident count. This is reflected in the regression results discussed in the next section.

Fourth, the change in the proportion of serious incidents relative to total or maneuvering incidents before and after the agreement suggests that either or both of two factors may be at work: an absolute decrease in the likelihood of serious incidents, and a change in incident reporting standards. The former would clearly decrease the proportion by directly reducing the numerator (and to a much smaller degree, the denominator); the latter would increase the denominator. Both effects would result in a decrease in the ratio between serious and non-serious incidents. The change in incident reporting standards again raises the issue of reporting bias. I will return to the question of reporting bias after evaluating the regression results.

**Interpreting the Regression Results**

As the focus in this analysis is on the effect of the INCSEA Agreement, I confine the discussion here to the relevant independent variable (\texttt{agree}: the presence of the agreement).\footnote{The regression results are in the appendix; a full discussion of the control variables and methodology are in Ling 2012.} The hypothesized effect of \texttt{AGREE} was negative: the agreement should reduce the incident rate.

The significance and direction of the \texttt{AGREE} coefficient is a challenge to
interpret, because the results are highly variable:

- Only one specification had a clearly significant negative coefficient for AGREE: the MANINC negative binomial specification with the LOG(TIME) control. In this specification, the LOG(TIME) coefficient was positive and significant. The conclusion is that this specification showed the agreement reducing the probability of an incident, but that this effect weakened over time. Given the relative sizes of the coefficients and the slow rate of increase in the LOG(TIME), their combined effect was strongly negative throughout the 15 post-agreement years.\(^66\)

- Specifications without the LOG(TIME) control all had positive (but marginally or not significant) AGREE coefficients.\(^67\)

- The final specification (MANINC hurdle with LOG(TIME) control) had a negative but insignificant AGREE coefficient.\(^68\)

The variation in the AGREE coefficient results is hard to interpret. The null results of no difference before and after the agreement are fairly straightforward to explain: the agreement had no effect. Similarly, the significant negative AGREE coefficients in the LOG(TIME) specifications also have a clear interpretation: the agreement did reduce incidents as intended.

The (almost significant) positive AGREE coefficients are harder to interpret. If we credit the coefficient sign, did the agreement actually make things

\(^{66}\)log(15) = 2.71. Applying this, the combined effect in 1988 would have been \(-4.8 + (2.71 \times 0.55) = -3.3\).

\(^{67}\)The p-values for the ALLINC and MANINC hurdle models were 0.056 and 0.0503 respectively.

\(^{68}\)The p-value was 0.078.
worse? This seems unlikely given the qualitative results. There is no plausible theoretical explanation for how a CBM agreement with readily observable standards for evaluating mutual compliance could encourage rather than discourage maneuvering incident behavior (especially when serious incidents dramatically decreased). The most likely explanation for this counterintuitive empirical result is reporting bias: individual maneuvering incidents were less likely to have been reported prior to the agreement. The next section discusses the evidence for and potential impact of reporting bias.

Evaluating Reporting Bias in the INCSEA Agreement

Consider an anecdotal account of an event that occurred in the Mediterranean Sea in the mid-1980s: a Soviet destroyer performing a close surveillance of a U.S. ammunition ship pointed its guns at the American vessel.\textsuperscript{69} The incident was not recorded in the central U.S. Navy INCSEA archives and thus is not part of the incident dataset used in this paper. Why was this apparent INCSEA violation not recorded?

One possibility is that the U.S. ship commander may not have reported it, either judging it unworthy of higher-level attention or because of concern that it would reflect poorly on him.\textsuperscript{70} Another possibility is that more senior commanders up the chain may have disregarded or refused to forward a submitted report, considering the incident insufficiently serious, inadequately

\textsuperscript{69}This example is based on the personal experience of Dr. David Winkler. Winkler oral interview with author, March 19, 2012.

\textsuperscript{70}Tamuz 1987, p. 70 discusses these factors in the context of aviation safety incident reporting. Winkler pointed out that this particular incident was not a credible threat. The Soviet ship was jeopardizing its own safety if it had seriously harbored hostile intent: had the guns been fired, the explosion of the American ammunition ship would have heavily damaged the Soviet vessel as well.
documented, or potentially embarrassing.

As discussed earlier, reporting error is likely to understate the true prevalence of incidents: not all incidents will realistically be reported. Bias in the other direction (overstating incidents) is less likely because the typical documentation required for INCSEA incident reporting would tend to make this difficult.\footnote{Allen 1990, p. 44 discusses the thorough documentation of incident reporting.} However, there are possible strategic incentives for over-reporting: as one U.S. INCSEA negotiator recalled of the early post-agreement period, “neither side wanted the other to have more claims.” Also, complaints could be raised in support of continuing negotiation interests.\footnote{Author interview with Rear Admiral Ronald J. Kurth, U.S. Navy (retired), May 31, 2012.}

Serious incidents are less likely to be subject to reporting bias, as Levitt pointed out for crime reporting: “murder, in contrast to other crimes, is likely to be immune from reporting bias since virtually all murders are likely to be reported to the police.”\footnote{Levitt 1999, p. 72.} Collision or fatal accidents are similar to murder in the naval context: both will almost certainly be immediately reported. Less serious incidents are more likely to be subject to reporting bias because of the greater possibility of doubt over the event’s true significance or the greater ability to conceal the incident from the chain of command.

The specificity of the INCSEA agreement is a probable source of bias. Several provisions of the agreement – particularly the prohibition of simulated attacks – were entirely new obligations, and thus went largely undocumented in the pre-agreement incident records. Also, the implementation of the agreement provided both sides with incentives to create more efficient and systematic internal reporting processes.
Consider two possible ways that the INCSEA Agreement could have reduced dangerous naval activity. First, it might have reduced just the likelihood of an incident escalating to a serious accident (presumably by limiting only the most dangerous behavior) while leaving the likelihood of less-serious incidents unaffected. Second, the agreement may have equally affected both the baseline probability of an incident occurring (Pr(Incident), labeled $p$ hereafter) and also the conditional probability of an incident escalating to serious status (Pr(Serious | Incident), labeled $s$ hereafter). In the former case, the ratio between total and serious incidents would change; in the latter, the ratio would remain unchanged.\textsuperscript{74} Both of these represent changes in the underlying baseline activity – not of reported activity, which is subject to possible bias.

Reporting bias is thus a potentially important effect on the probability of capturing non-serious incidents in the INCSEA dataset. Its effects will be larger before the agreement for a range of plausible assumptions. This will in turn cause the apparent effectiveness of the agreement in reducing incidents to be understated, as the regression results suggest.

**Quantifying Possible INCSEA Reporting Bias**

Quantifying the impact of reporting bias is a challenging problem. Levitt’s technique of examining the ratio between murder and lesser crimes is relevant to the INCSEA problem. Murder is assumed to always be reported, while the reporting of lesser crimes such as robbery or assault is subject to bias.

\textsuperscript{74}Of course, an intermediate possibility where both probabilities change but the serious incident probability changes more is more likely, but the no-change and identical-change options bound the problem.
If the actual ratio between murder and nonmurder crimes is considered to be constant, then changes in the ratio of murder to lesser crime reports can provide information on the extent of reporting bias (or at least bound the estimates of bias).\(^75\)

Turning to the INCSEA data, the ratio of reported maneuvering incidents to serious incidents increased over four-fold after the agreement was signed.\(^76\) If we assume that collisions and fatalities were always reported, then the true population proportion of incidents that escalated into serious incidents declined from 0.047 to 0.011.

The two hypotheses presented earlier both suggest that one of the relevant baseline incident probabilities (the conditional probability of an incident escalating to serious incident status \(s\) for \(H1\), or the conditional probability of an incident occurring \(p\) for \(H2\)) decreased after the INCSEA Agreement was signed. Given one reasonable assumption,\(^77\) the empirical data suggest that at least one (and possibly both) of the hypotheses must be correct.

To evaluate this, we assume in turn that each of the hypotheses is wrong, and evaluate what the reporting bias implications of those assumption are. First, we assume that the actual serious-incident conditional probability \(s\) is unchanged from before to after the INCSEA agreement. If this is true, then pre-agreement reporting is biased low and the pre-agreement baseline incident probability \(p_1\) is twice as high as the post-agreement probability \(p_2\).

\(^{75}\)Levitt 1998, pp. 71-76.

\(^{76}\)The actual figures are 21.3:1 before 1972 and 87:1 after 1972. The comparable reported total-incident to serious incident ratios were 23:1 and 150:1. Note that the serious incidents are included in the both the numerator and denominator of these ratios.

\(^{77}\)Specifically, that all serious incidents were reported both before and after the agreement.
Second, if we assume that the actual baseline incident probability $p$ remains unchanged before and after the agreement, then the post-agreement reporting rate is biased low and the pre-agreement serious incident conditional probability $s_1$ is roughly 3.5 times greater than the post-agreement probability $s_2$. For an intermediate case in which the pre-agreement non-serious incident reporting rate is half of the post-agreement rate, $s_1 > s_2$ and $p_1 > p_2$. In this case, both hypotheses are correct.\footnote{Proofs for these assertions are in the Appendix.}

5.0.1 Assessing INCSEA Effectiveness: A Combined Cut

Qualitative assessments of the INCSEA agreement have been almost uniformly positive, arguing that it both provided workable behavioral standards and a very effective consultative mechanism. The quantitative analysis provides qualified support for the qualitative literature’s conclusions.

The post-INCSEA record shows a dramatic decrease in the serious incident rate and a more moderate decrease in the rate of maneuvering incidents. The total-incident rate increases after the agreement, but this is driven by a seven-fold increase in non-maneuvering incidents. This indicates that the maneuvering incident count is the most reliable dependent variable.

The reporting bias analysis provides a credible explanation for the observed increase in the reported all-incident and non-maneuvering incident rates. The INCSEA Agreement created incentives and processes for improved reporting that were absent pre-agreement. This in turn would make the pre-agreement incident counts artificially low, biasing the event count regression results and understating the agreement’s true effectiveness in re-
ducing incident behavior.

6 Conclusion

Reporting bias presents a potentially serious obstacle for inference in assessing the effectiveness of international law. Systematic variation in the probability of recording data will cause some data (events or counts) to be missing. Missing events will often vary from non-missing data in inferentially important ways. The missing data will bias estimates of an agreement’s compliance or effectiveness.

Dealing with reporting bias is challenging because we often do not (and sometimes realistically cannot) know the scope of the missing data problem. Estimating the possible impact of reporting bias is difficult when we cannot determine the size of the reportable event population.

The multiple potential causes of reporting bias – observability, clarity, reporting incentives, and capacity – provide insights into how to identify the presence of reporting bias. They also suggest possible ways to improve future reporting systems or to re-examine existing data to offset biased reporting.

The effects of reporting bias vary according to – along with other, domain-specific factors – the nature of the data, the type of variable, and the source of the bias. If international law scholars can trace probable sources and causes of reporting bias, they will be better able to produce qualitative or quantitative estimates of the amount of bias in their dependent variables and thus gain a better sense for the uncertainty in their analyses.

Finally, in some cases the effects of reporting bias can be minimized or
mitigated. Exploitation of events that are always reported — such as the severe naval incidents in the INCSEA case — can provide a critical foundation of certainty that enables better identification and bounding of uncertainty elsewhere in the data. Application of an expected population distribution — such as the power law in Friedman’s analysis of casualties in the Indian Wars — can provide a credible estimate of both the extent of missingness and of the likely values of the missing data itself. And as several of the examples show, there is often no substitute for comparing multiple sources of information to identify biases in reporting processes.

Evaluating the causal effect of international law has always been a challenging inferential problem. The classical endogeneity problem is not the only challenge, however. Recognizing that the reporting bias challenge exists can enable scholars to identify, bound, and possibly mitigate its negative effects on the effort to understand whether and how international law can affect state behavior.
### Data and Analytical Appendix

### Statistical Results

**Table 1: Variables and Summary Statistics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>INCIDENTS(ALL)</th>
<th>INCIDENTS(MANEUV)</th>
<th>WEIS</th>
<th>FLEETSIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Incidents (All types) per month</td>
<td>Incidents (Maneuvering) per month</td>
<td>WEIS scaled score</td>
<td>U.S.-Soviet Fleet Size</td>
</tr>
<tr>
<td><strong>AGREE = 0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>-98.0</td>
<td>489</td>
</tr>
<tr>
<td>Maximum</td>
<td>11</td>
<td>9</td>
<td>36.7</td>
<td>583</td>
</tr>
<tr>
<td>Mean</td>
<td>1.09</td>
<td>0.99</td>
<td>-16.6</td>
<td>539</td>
</tr>
<tr>
<td>Std error</td>
<td>(1.93)</td>
<td>(1.72)</td>
<td>(20.0)</td>
<td>(25.7)</td>
</tr>
<tr>
<td><strong>AGREE = 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>-304.7</td>
<td>446</td>
</tr>
<tr>
<td>Maximum</td>
<td>14</td>
<td>10</td>
<td>77.5</td>
<td>520</td>
</tr>
<tr>
<td>Mean</td>
<td>1.51</td>
<td>0.88</td>
<td>-22.8</td>
<td>482</td>
</tr>
<tr>
<td>Std error</td>
<td>(2.21)</td>
<td>(1.56)</td>
<td>(38.9)</td>
<td>(20.1)</td>
</tr>
<tr>
<td><strong>Full series</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>-304.7</td>
<td>446</td>
</tr>
<tr>
<td>Maximum</td>
<td>14</td>
<td>10</td>
<td>77.5</td>
<td>583</td>
</tr>
<tr>
<td>Mean</td>
<td>1.33</td>
<td>0.93</td>
<td>-20.2</td>
<td>506</td>
</tr>
<tr>
<td>Std error</td>
<td>(2.10)</td>
<td>(1.63)</td>
<td>(32.3)</td>
<td>(36.4)</td>
</tr>
</tbody>
</table>
Table 2: Monthly Incident Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Pre-INCSEA</th>
<th>Post-INCSEA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Months</td>
<td>149</td>
<td>199</td>
<td>348</td>
</tr>
<tr>
<td>Σ Serious Incidents</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Mean (Serious Incidents)</td>
<td>0.047</td>
<td>0.010</td>
<td>0.026</td>
</tr>
<tr>
<td>Std Dev (Serious Incidents)</td>
<td>0.24</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td>Σ All Incidents</td>
<td>163</td>
<td>300</td>
<td>463</td>
</tr>
<tr>
<td>Mean (All Incidents)</td>
<td>1.1</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Std Dev (All Incidents)</td>
<td>1.9</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Σ Maneuvering Incidents</td>
<td>149</td>
<td>174</td>
<td>323</td>
</tr>
<tr>
<td>Mean (Maneuvering Incidents)</td>
<td>1.0</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>Std Dev (Maneuvering Incidents)</td>
<td>1.7</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Σ Non-Maneuvering Incidents</td>
<td>14</td>
<td>126</td>
<td>140</td>
</tr>
<tr>
<td>Mean (Non-Maneuvering Incidents)</td>
<td>0.094</td>
<td>0.63</td>
<td>0.40</td>
</tr>
<tr>
<td>Std Dev (Non-Maneuvering Incidents)</td>
<td>0.39</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Probability of 0 All-Incident Month</td>
<td>0.57</td>
<td>0.41</td>
<td>0.48</td>
</tr>
<tr>
<td>Probability of 0 Maneuvering-Ionicent Month</td>
<td>0.58</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td>Ratio Serious:Maneuvering Incident</td>
<td>1:21.3</td>
<td>1:87</td>
<td>1:35.9</td>
</tr>
<tr>
<td>Prob(Serious</td>
<td>Maneuvering Incident)</td>
<td>0.047</td>
<td>0.011</td>
</tr>
<tr>
<td>Ratio Serious:All</td>
<td>1:23.3</td>
<td>1:150</td>
<td>1:51.4</td>
</tr>
</tbody>
</table>
Table 3: $\chi^2$ Test of Independence, Pre-Post Agreement

<table>
<thead>
<tr>
<th>Comparison</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious-Maneuvering Incidents</td>
<td>2.54</td>
<td>1</td>
<td>0.119</td>
</tr>
<tr>
<td>Serious-Total Incidents</td>
<td>5.51</td>
<td>1</td>
<td>0.0189</td>
</tr>
</tbody>
</table>

Table 4: Incidents (All) - Monthly Count Distribution

<table>
<thead>
<tr>
<th>Count</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>169</td>
<td>0.486</td>
</tr>
<tr>
<td>1</td>
<td>74</td>
<td>0.213</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>0.135</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>0.060</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>0.037</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.014</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.017</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.011</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.009</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.006</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: Incidents (Manuevering) - Monthly Count Distribution

<table>
<thead>
<tr>
<th>Count</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>201</td>
<td>0.622</td>
</tr>
<tr>
<td>1</td>
<td>76</td>
<td>0.235</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>0.102</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0.050</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.015</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.015</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0.012</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.012</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.006</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 6: INCSEA Regression Results

<table>
<thead>
<tr>
<th>Method</th>
<th>NegBin ALLInc</th>
<th>Hurdle ALLInc</th>
<th>NegBin MANInc</th>
<th>Hurdle MANInc</th>
<th>NegBin MANInc</th>
<th>Hurdle MANInc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.12</td>
<td>-1.7</td>
<td>-5.1*</td>
<td>-8.2**</td>
<td>-0.019</td>
<td>-4.0</td>
</tr>
<tr>
<td></td>
<td>(2.1)</td>
<td>(2.9)</td>
<td>(2.4)</td>
<td>(3.2)</td>
<td>(3.0)</td>
<td>(3.7)</td>
</tr>
<tr>
<td>WEIS</td>
<td>-0.0079*</td>
<td>-0.010*</td>
<td>-0.011**</td>
<td>-0.013**</td>
<td>-0.0068</td>
<td>-0.0083</td>
</tr>
<tr>
<td></td>
<td>(0.0032)</td>
<td>(0.0047)</td>
<td>(0.0035)</td>
<td>(0.0046)</td>
<td>(0.0038)</td>
<td>(0.0049)</td>
</tr>
<tr>
<td>FLEETSIZE</td>
<td>-0.00030</td>
<td>0.0022</td>
<td>0.0092*</td>
<td>0.014*</td>
<td>0.0068</td>
<td>0.013*</td>
</tr>
<tr>
<td></td>
<td>(0.0039)</td>
<td>(0.0055)</td>
<td>(0.0044)</td>
<td>(0.0059)</td>
<td>(0.0045)</td>
<td>(0.0060)</td>
</tr>
<tr>
<td>AGREE</td>
<td>0.22</td>
<td>0.72</td>
<td>0.26</td>
<td>0.78</td>
<td>-4.8**</td>
<td>-3.6</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.38)</td>
<td>(0.31)</td>
<td>(0.40)</td>
<td>(1.8)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>LOG(TIME)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.55**</td>
<td>0.49*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.20)</td>
<td>(0.22)</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>—</td>
<td>2.7</td>
<td>—</td>
<td>2.4</td>
<td>—</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4.2)</td>
<td></td>
<td>(6.5)</td>
<td></td>
<td>(8.2)</td>
</tr>
<tr>
<td>WEIS</td>
<td>—</td>
<td>-0.0060</td>
<td>—</td>
<td>-0.0075</td>
<td>—</td>
<td>-0.0046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0045)</td>
<td></td>
<td>(0.0056)</td>
<td></td>
<td>(0.0056)</td>
</tr>
<tr>
<td>FLEETSIZE</td>
<td>—</td>
<td>-0.0049</td>
<td>—</td>
<td>-0.0055</td>
<td>—</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0075)</td>
<td></td>
<td>(0.011)</td>
<td></td>
<td>(0.012)</td>
</tr>
<tr>
<td>AGREE</td>
<td>—</td>
<td>-0.40</td>
<td>—</td>
<td>-0.79</td>
<td>—</td>
<td>-8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.51)</td>
<td></td>
<td>(0.72)</td>
<td></td>
<td>(4.2)</td>
</tr>
<tr>
<td>LOG(TIME)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.43)</td>
</tr>
</tbody>
</table>

Predicted Zero: 170 169 203 201 203 201
Actual Zero: 169 169 201 201 201 201
Reporting Bias Analysis

Define $p$ as the baseline monthly probability that an incident will occur, and $t$ as the number of months in the analysis period. Define $s$ as the probability that an incident, if one occurs, will escalate to become a serious incident, and $r$ as the probability that an incident, if non-serious, will be reported. Assign the subscript 1 for pre-agreement and 2 for post-agreement. We thus have:

\[
\begin{align*}
    p &= \text{Prob(Incident)} \\
    s &= \text{Prob(Serious | Incident)} \\
    r &= \text{Prob(Non-serious Incident is reported)} \\
    t &= \text{Time (number of months)}
\end{align*}
\]

Hypothesis 1 can thus be summarized as $s_2 < s_1$, and hypothesis 2 as $p_2 < p_1$. However, we cannot directly observe $p$, $r$, or $s$.

The observed count of serious incidents can be expressed, however, as $p s t$. (We assume that the probability of reporting a serious incident is one.) The observed total count of incidents can be expressed as $p r t$.

We now define a new term $b$ for reporting bias:

\[
b = \frac{r_2}{r_1}
\]

Assuming that reporting rates are higher after the INCSEA Agreement than before, $r_2 > r_1$ and therefore $b > 1$.

\[79\text{Strictly speaking, the observed total count is } p r t + p s t, \text{ but we assume that the number of serious accidents is small compared to the total number and therefore that } p r t + p s t \approx p r t.\]
Using observed serious and maneuvering incident rates and maintaining the assumption that \( s << p \), the pre-agreement observed ratio of serious to all maneuvering incidents is:

\[
\frac{\text{observed serious incidents}}{\text{observed maneuvering incidents}} = \frac{7}{149} = \frac{s_1 p_1 t_1}{r_1 p_1 t_1} = \frac{s_1}{r_1} = 0.0470
\]

with the similar post-agreement ratio as

\[
\frac{\text{observed serious incidents}}{\text{observed maneuvering incidents}} = \frac{2}{174} = \frac{s_2 p_2 t_2}{r_2 p_2 t_2} = \frac{s_2}{r_2} = 0.0115
\]

We thus observe that \( \frac{s_1}{r_1} = 4.08 \frac{s_2}{r_2} \) for maneuvering incidents.

We first assume that Hypothesis 1 is wrong, and that there is no change in the conditional serious incident rate. Therefore, \( s_1 = s_2 \). If we accept this statement as true, beginning with the observed serious incident counts:

\[
\frac{s_1}{r_1} = 4.08 \frac{s_2}{r_2} \\
\frac{s_2}{r_1} = 4.08 \frac{s_2}{r_2} \\
\frac{1}{r_1} = 4.08 \frac{1}{r_2} \\
\frac{r_2}{r_1} = 4.08 = b
\]

Thus, \( r_2 > r_1 \) and \( b > 1 \). If the conditional serious incident rate is unchanged, then non-serious incidents are 4.08 times more likely to be reported after the INCSEA agreement than before the agreement.

The observed maneuvering incident counts (\( r_1 p_1 t_1 \) and \( r_2 p_2 t_2 \)) are 149 and 174 respectively. The before/after ratio of observed maneuvering incidents
reports is therefore $149/174 = 0.86$.

$$r_1p_1t_1 = 0.86r_2p_2t_2$$

$$p_1 = 0.86 \frac{r_2}{r_1} t_2 p_2$$

$$p_1 = 0.86 b \frac{t_2}{t_1} p_2$$

$$p_1 = 0.86(4.08) \frac{199}{149} p_2$$

$$p_1 = 2.11 p_2$$

Given the reporting bias $b$ calculated from the assumption that H1 is false and that the true conditional serious incident probability is unchanged ($s_1 = s_2$), the observed data indicate that the underlying incident probability $p_1 > p_2$, supporting the H2 claim that the baseline maneuvering incident rate was higher before the agreement rather than after.\(^\text{80}\)

Now, assume that H2 is false (implying that the INCSEA Agreement had no effect on the baseline maneuvering incident probability with $p_1 = p_2$).

$$r_1p_1t_1 = 0.86r_2p_2t_2$$

$$r_1p_1t_1 = 0.86r_2p_1t_2$$

$$r_1t_1 = 0.86r_2t_2$$

$$\frac{t_1}{0.86t_2} = \frac{r_2}{r_1} = b$$

$$0.87 = b$$

\(^{80}\)Replacing the maneuvering incident rate with the total incident rate and performing a similar calculation results in $p_1 = 4.66p_2$, also supporting the H2 prediction that $p_1 > p_2$.\(^\text{52}\)
Applying this reporting bias to the observed pre- and post-agreement ratios of serious to reported maneuvering incidents:

\[
\frac{s_1}{r_1} = 4.08 \frac{s_2}{r_2} \\
s_1 = 4.08 \frac{r_2}{r_1} s_2 \\
s_1 = 4.08 b s_2 \\
s_1 = 4.08 (0.87) s_2 \\
s_1 = 3.55 s_2
\]

Thus, if baseline maneuvering incident probability is assumed to be unchanged, then the serious incident conditional probability decreased by a factor of 3.5 and H1 is confirmed.\textsuperscript{81}

Finally, with reasonable assumptions for \( b \), both H1 and H2 can simultaneously be true. Assume that \( b = 2 \): pre-agreement reporting is half as efficient as post-agreement reporting. In this case, calculations similar to those above will demonstrate that \( p_1 = 1.72 p_2 \) and \( s_1 = 2.04 s_2 \), supporting both H1 and H2.

\textsuperscript{81}Repeating the calculations using the total incident count instead of maneuvering incidents produces the result \( s_1 = 8.8 s_2 \), also confirming H1’s assertion that \( s_1 > s_2 \).
Works Cited


Tyldum, Guri and Anette Brunovskis. 2005. “Describing the Unob-


