The “Dirty War Index” and the Real World of Armed Conflict

Amelia Hoover, Romesh Silva, Tamy Guberek, and Daniel Guzmán

May 23, 2009

1 Introduction

Is the crisis in Darfur a genocide? How common is sexual violence in the Democratic Republic of Congo? Is it “widespread and systematic” tactic of one or more armed groups? Did Colombia’s paramilitary demobilization program actually reduce violence against noncombatants? How many Iraqi civilians have died during the ongoing American-led conflict there? How does that number compare to the number of soldiers and insurgents killed?

Each of these high-stakes questions asks the analyst to estimate the “dirtiness” of a military action. Directly or indirectly, each can be (and has been) answered using quantitative evidence. Different answers point to radically different legal and policy conclusions.

Given the dire consequences of wrong answers, quantitative research in human rights demands accuracy—or, at the least, an honest accounting of uncertainty. Yet armed conflicts (like many other human rights crises) are near-impossible contexts for the collection of comprehensive, or even representative, data. Simple statistical methods that nevertheless delivered accurate results could revolutionize human rights research.

The Dirty War Index (DWI) [1], a measure recently proposed by Hicks and Spagat, is indeed simple. It does promise to facilitate quantitative analysis of factual disputes in international humanitarian law. But the authors’ claims to accuracy—in particular, the statement that ratios such as the DWI are “relatively less affected by under- or over-counting than absolute numbers,” [1:1662] and that accurate DWI calculations can be produced from “any data source” [1:1662]—do not withstand scrutiny.

In this article, after briefly reviewing the definition of the DWI and its claimed advantages, we offer two related critiques. First, we demonstrate that accuracy claims about the DWI depend on the unmet assumption that rates of over- or under-counting (sampling bias) are equal across space, time, armed group and other dimensions. Second, we demonstrate the unreliability of various DWI measures, using data from El Salvador, Guatemala, Colombia and East Timor. We conclude by outlining best practices for human rights data collection and analysis.

2 About the DWI

Hicks and Spagat define the DWI as a simple rate: the number of "dirty" cases or outcomes, A, divided by the total number of outcomes, $A+B$, multiplied by 100. [1: 1658]

$$DWI = \frac{A}{A+B} \times 100 \tag{1}$$

For example, an armed group that caused 10,000 fatalities, of which all were recorded and 5,000 were illegal ("dirty") noncombatant fatalities, would have a noncombatant-fatality DWI of $\frac{5{,}000}{10{,}000} \times 100 = 50$. Drawing on event data from the Colombian civil war, Hicks and Spagat calculate a number of DWI measures and identify the international legal statutes to which they apply.

Hicks and Spagat argue that use of the DWI statistic has a number of practical and theoretical benefits. While raw event data captures the *magnitude* of undesirable outcomes, Hicks and Spagat argue that the *proportion* of dirty outcomes is a more valuable statistic because proportions are "explicitly linked to international humanitarian law" [1:1658]. Additional Protocol 1 to the Fourth Geneva Convention of 1949, for example, prohibits among other things "incidental civilian casualties disproportionate to the advantage gained in attacking a military target" [2].

No other international legal standard references proportionality; still, the authors argue that proportions—of captured combatants executed, of noncombatant women sexually assaulted, and of combatants who are children, among others—are more valuable than absolute magnitudes because they "lend themselves to comparisons over time, between wars, between weapons, and between warring combatant groups to identify better versus worse performers" [1:1658].

Furthermore, claim Hicks and Spagat, DWI statistics can be computed from "any data source (media reports, epidemiological surveys, coroners' reports)" [1:1661], and "can be easily used and understood, facilitating interdisciplinary communication and research on war's effects" [1:1659].

Each of the DWI's advantages presupposes the accuracy of the measure. As we discuss below, however, its accuracy depends upon assumptions about over- and under-counting that are seldom, if ever, met in a conflict situation.

3 The Equal Under-Counting Assumption

Hicks and Spagat state that "[a]lthough bias can affect DWI values, as ratios DWIs are relatively less affected by under- or over-counting than absolute numbers. For example, if a population generally under-reports war-related rape by 40%, this does not bias comparing rates between different combatant groups" [1:1662].

This statement lays bare a key assumption of the DWI enterprise, which we call the "equal under-counting assumption." Throughout their work, Hicks and Spagat assume that, except in cases of outright manipulation of data, rates of under-registration[1] are very similar or identical across different locations,

[1] Note that we use the terms "under-registration", "under-reporting" and "under-counting" interchangeably—and that we focus on under-counting because it is significantly more common than over-counting.

perpetrator groups, victim groups, violation types and other dimensions of a conflict.

The equal under-counting assumption can be expressed mathematically with some simple algebra: for true dirty outcomes A, true non-prohibited outcomes B, and under-counting multipliers i, j, if

$$\widehat{DWI} = DWI \tag{2}$$

then the following equality must hold:

$$\frac{A}{A+B} = \frac{\hat{A}}{\hat{A}+\hat{B}} = \frac{Ai}{Ai+Bj} \tag{3}$$

i.e.,

$$i = j. \tag{4}$$

Consider the hypothetical scenario above in Section 2, in which the true number of noncombatant deaths is 5,000 and the true number of combatant deaths is 5,000 (for a combined total of 10,000 deaths). If clean and dirty outcomes are reported at an equal rate (say, 40%), the measured DWI is equal to the true DWI, as Hicks and Spagat suggest: $\frac{0.4\times5{,}000}{0.4\times5{,}000+0.4\times5{,}000} \times 100 = 50$.

But what if clean outcomes are reported more frequently than dirty outcomes (as is very common)? If $i \neq j$, then the DWI may be very inaccurate indeed. If only 25% of noncombatant fatalities are recorded while 75% of combatant fatalities are recorded, the DWI is $\frac{0.25\times5{,}000}{0.25\times5{,}000+0.75\times5{,}000} \times 100 = 25$, half the true proportion of dirty outcomes.

To take a less abstract example, Hicks and Spagat report DWI statistics for Colombian armed groups, assuming equal levels of under-reporting across six very different categories of violence: civilians killed by paramilitaries, civilians killed by guerrillas, civilians killed by government forces, combatants killed by paramilitaries, combatants killed by guerrillas, and combatants killed by government forces [1:1659]. The greater the difference between reporting rates for civilian deaths and reporting rates for combatant deaths, the more misleading an individual DWI will be; if reporting also varies across fatalities caused by different forces (government, paramilitary and guerrilla), then comparisons between the groups will be similarly misleading.

The accuracy and comparability of DWI statistics depend fundamentally upon the equal under-counting assumption—yet the assumption is entirely implausible in a conflict situation. Reporting rates for violent incidents vary in complex and unpredictable ways within any given conflict: by region, by perpetrator, and by key victim characteristics such as gender, race, ethnicity and combatant status. This variation typically results not from malfeasance but from complexity and scarcity in the reporting process. Social networks, trust, resource availability, grant-making, access to violent areas, conflict-related migration flows, and even weather can strongly—but not randomly—affect reporting rates at the micro level.

Perhaps the most analytically frustrating source of bias in violence statistics is violence itself: violent areas or situations are often inaccessible to human rights observers, government officials or other monitors, such that particularly extreme outbursts of violence may be severely underreported until well after the fact. Areas of extreme violence are likely to have high proportions of "dirty" outcomes, and unlikely to be available for measurement. In this situation, DWI statistics would inflate the proportion of "clean" outcomes, obscuring what they are intended to measure.

This is a practical, not a solely theoretical critique. As we illustrate in Section 5 below, extremely uneven under-reporting of violent events is the rule, not the exception.

4 DWIs and Common Data Sources

Hicks and Spagat state that the DWI can be calculated using "any type of data" [1:1661]. However, given the sensitivity of the DWI measure to within-case variation in under-registration, the condition that data be "adequately valid, accurate and comprehensive" effectively disqualifies most real-world data sources. In our experience, datasets that meet Hicks' and Spagat's conditions for "adequacy" are exceedingly rare.[2][4][5][6][7] In this section we discuss in turn three types of datasets that the authors suggest will be "adequately valid, accurate and comprehensive": epidemiological surveys, media reports and coroners' reports. We conclude, *contra* Hicks' and Spagat's assertion, that DWIs can seldom be confidently calculated from any of these sources.

Like most surveys, epidemiological surveys are based on probability sampling: a random sample is drawn from the population in order to measure the rate of occurrence of a phenomenon within that population. Unfortunately, probability samples typically perform relatively poorly as measures of elusive phenomena [8][9]—such as direct conflict mortality. For example, sexual violence, killings, forced disappearances and arbitrary detentions will likely not be realiably captured by a random, population-based survey. Even a carefully designed cluster sample survey of well-remembered rare events such as deaths, recalled over a short period, tends to produce under-estimates with very wide confidence intervals.

In their work estimating noncombatant mortality in Kosovo, Spiegel and Salama [10] report "a number of limitations to our study," each of which "would have led to an underestimation of mortality but would not have altered our major conclusions. [T]he small number of deaths and missing people in our study has resulted in wide [confidence intervals]." Summarizing, Spiegel and Salama maintain that "Current population-based methodologies, such as cluster surveys, require technical expertise and considerable logistic support and may lack precision when documenting rare events such as deaths."

Given that high-quality epidemiological surveys produce wide confidence intervals even for estimates of easily identifiable phenomena, such as conflict mortality, it would be inappropriate to conduct calculations like the DWI using simple point estimates. We also note that confidence intervals for subsets of conflict mortality statistics, such as "dirty" versus "clean" mortality, are wider still than those for global estimates. Hence, the ability of the DWI, applied to survey-based data, to identify proportional differences with the required statistical power is necessarily weak.

Passive surveillance systems, such as media reports, hospital records, and coroners' reports, when assessed in isolation, lack even the benefit of a defen-

[2] Bosnian Book of the Dead, a near-comprehensive account of killings during the civil war in Bosnia, is the single exception to this rule known to the authors.[3]

sible confidence interval. Like most data about violence, these are convenience samples—non-random subsets of the true universe of violations. Any convenience dataset will include only a fraction of the total cases of violence; as discussed in the preceding section, fractional samples can never be assumed to be random or representative.

Even in industrialized countries with well-developed administrative systems, such passive surveillance systems are vulnerable to errors and bias. Monkonnen demonstrated that the proportion of homicides recorded in newspapers and coroners' varied over time.[11] Monkkonen further noted that both coroners' reports and media reports systematically underestimated homicides in years of high homicide incidence.

DWIs cannot with confidence be constructed from any single uncorrected data source. While the quality of the underlying estimates (of dirty and clean outcomes) varies widely between epidemiological surveys and convenience samples, even the best single data source is almost always insufficient. Although the process is necessarily more complicated, drawing from and triangulating between several data sources often provides the most accurate portrait of data collection processes and (by implication) the violent events those processes were intended to measure.

An approach which encompasses multiple data sources and complementary measurement strategies is required to clarify both absolute and relative magnitudes of different violent outcomes. In Section 5.3 below, drawing on recent findings in Timor-Leste, we show how a valid decomposition of direct versus indirect conflict mortality in protracted conflicts requires multiple, complementary data sources and methods to overcome the significant inherent flaws of individual data sources and methods.

5 Examples

In this section, we present data excerpts from a number of conflicts and discuss their implications for the utility of the DWI. Our analysis focuses on the comparison of multiple, independently-collected samples. We show that across samples, there is generally a wide variation in reporting rates over time, space, and perpetrator. Even small differences in true levels of dirty outcomes across any of these dimensions may be strongly amplified (or entirely obscured) by reporting rate variations.

In Guatemala, press sources, in particular, completely failed to capture the massive rise in violence that accompanied the Guatemalan Army's genocidal *tierra arrasada* campaign across the Mayan highlands in the early 1980s. Each data source on the Guatemalan conflict, taken in isolation, leads to fundamental misunderstandings of the overall conflict. Similarly, three major data sources on El Salvador's 1980-1992 conflict cover "dirty" outcomes at rates that vary substantially over both space and perpetrating group. Again, choosing any individual data source, or comparing DWI statistics from different data sources, would lead to serious misinterpretations of intra-conflict dynamics.

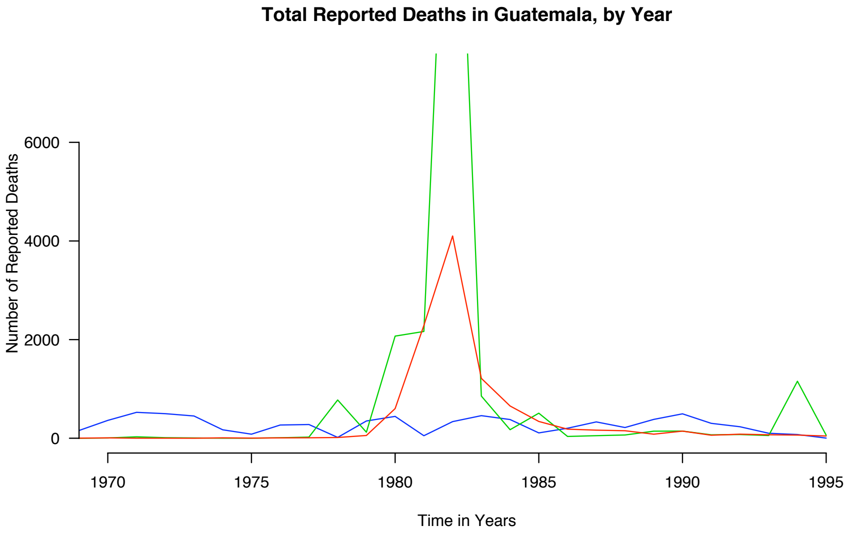
Our third example considers data from Timor-Leste, comparing a convenience sample of narratives from an oral history project with a household mortality survey. We show that although the two sources are superficially similar, they yield DWI's differing by more than a factor of three. We close by discussing how deliberate data manipulation—in this example, the case of Colombia's "false guerrillas"—affects DWI measurements.

5.1 Guatemala: The Inadequacy of a Single Convenience Sample, I

During Guatemala's internal armed conflict (1960-1996), several data collection efforts documented ongoing violence. The Guatemalan case illustrates how different convenience samples, given differential access to geographic areas and social networks, can present an almost inverse account of the past.

The graph below shows killings and disappearances in Guatemala during the conflict, as recorded by three different sources: a comprehensive review of the press (blue line), documentation by non-governmental human rights sources (red line),[3] and testimonies by witnesses and victims' family members (green line). Most notably, media sources reported close to zero cases during the early 1980's, while retrospective testimony (and, to a lesser extent, human rights casefiles) show a massive spike in violence against noncombatants during these years. This is due to the concentration of media resources and attention in urban areas.

Documentary and interview sources, on the other hand, largely failed to capture data on low-level urban repression during the 1960s and 1970s, which *was* covered in the press. In addition, these sources reported only a fraction of urban violence in the years before and after its reported peak. The correlation between press data and all other sources for Guatemala is $\rho \approx 0$.[4]



[3] These include primarily the Guatemalan Human Rights Commission (CDHG), or in Spanish, *Comisión de Derechos Humanos de Guatemala* and the *Grupo de Apoyo Mutuo* (GAM).

[4] Other sources include interviews and documentary sources collected by the *Centro Internacional para Investigaciones en Derechos Humanos* (CIIDH), as well as the Catholic Church's Recovery of Historical Memory project (REMHI) and the UN Commission for Historical Clarification (CEH).

The report of the UN Commission for Historical Clarification (CEH) in Guatemala concluded that in early 1980's, the Guatemalan army committed acts of genocide against the Mayan population. [5] [12] In all, over 200,000 people died during three decades of conflict in Guatemala. Had the project depended exclusively on press sources, these facts would have remained utterly invisible.

The relevance of these findings to the DWI is clear. When an ostensibly reliable data source such as a comprehensive press database misses the vast majority of violence in a particular period, location, or group, the equal under-counting assumption is incorrect, and leads to incorrect conclusions. In the Guatemalan case, the vast majority of uncounted deaths were those of noncombatants—clearly a dirty outcome; yet a DWI statistic calculated from press data would have found the Guatemalan conflict relatively clean.

5.2 El Salvador: The Inadequacy of a Single Convenience Sample, II

The United Nations-sponsored Truth Commission for El Salvador (TRC) concluded that state and state-sponsored agents perpetrated the vast majority of human rights violations against noncombatants during that country's twelve-year civil war.[13] Beyond that basic finding, much of the intra-conflict variation in violence remains subject to debate. Indeed, even the number of persons killed in the conflict is subject to debate; non-scientific estimates range from 30,000 to 70,000 or more persons (i.e., approximately 0.6% of the pre-war population to approximately 1.4% of the pre-war population).

The three most comprehensive existing datasets on the Salvadoran conflict display large variations across space, perpetrating group, and other factors, an outcome that would suggest serious instability in DWI calculations even if total numbers of "clean" outcomes (violent outcomes occurring to combatants) could be accurately known. The list of violations against noncombatants reported to the TRC totals approximately 7,000, while the non-governmental Commission for Human Rights lists approximately 23,000 violations and the NGO El Rescate coded approximately 22,000 violations from the case files of the Archbishopric of San Salvador's Legal Aid office (Tutela Legal).

[5] The CEH's finding was supported by the multiple systems estimate that triangulated three sources to overcome their bias and coverage limitations.

Table 1: Percent of Violations by Department, Three Data Sources

Department	CDHES	Rescate	UNTRC
San Salvador	35.18%	30.60%	10.91%
Unknown	19.82%	5.11%	3.15%
San Miguel	6.85%	4.32%	1.46%
La Libertad	6.72%	6.24%	4.29%
Usulutan	6.70%	5.49%	3.05%
Morazan	5.12%	8.90%	6.35%
Ahuachapan	3.08%	0.67%	1.17%
Chalatenango	3.00%	7.41%	31.47%
La Paz	2.78%	3.11%	3.43%
Cuscatlan	2.66%	12.62%	11.93%
Santa Ana	2.61%	5.52%	2.98%
San Vicente	2.38%	5.38%	9.19%
Cabanas	1.44%	2.66%	7.90%
Sonsonate	1.25%	1.26%	2.29%
La Union	0.42%	0.71%	0.41%

In Table 1, note the large variations across El Salvador's fourteen departments, especially departments with high reporting density such as San Salvador and (in one case) Chalatenango. During the war, Chalatenango was a free-fire zone, highly lethal to its remaining civilian population and largely inaccessible to outside observers, while San Salvador, though less violent, was highly accessible. The types of violations reported in these three datasets are highly dependent upon social networks for the safe passage of persons and information; variations in reported levels of violence likely reflect variations in access or networks, rather than true spatial variations in the level of violence.

No reliable estimate of "dirty" (or, for that matter, clean) outcomes by department has yet been produced; this is consequential because perpetrating groups, including the five organizations of the FMLN (Frente Farabundo Martí para la Liberación Nacional) insurgent coalition and several state groups, tended to maintain geographically specific zones of control.

Table 2: Percent of Violations by Perpetrator Group, Three Data Sources

Perpetrator	CDHES	Rescate	UNTRC
Salvadoran Army	38.15%	47.29%	49.39%
National Police	15.62%	5.35%	1.53%
Death Squad	13.79%	5.55%	6.88%
Unknown	8.21%	19.31%	8.41%
National Guard	8.15%	3.52%	9.64%
Customs or Treasury Police	7.13%	2.04%	3.26%
Paramilitary	5.38%	2.85%	16.26%
Other	1.99%	6.78%	0.00%
FMLN	1.59%	7.30%	4.63%

In Table 2, note that data on perpetrating organizations, like data on geography, varies significantly by source. No accurate count of combatant deaths exists; however, the Truth and Reconciliation Commission lists approximately 11,000 armed forces deaths and approximately 1,100 FMLN deaths, with the caveat that the FMLN list is quite incomplete. Calculating DWI measures (noncombatants killed over total killings for each group) for each dataset based on these estimates of combatant deaths produces the following estimated DWI measures:

- $\widehat{CDHES}_{State} = \frac{0.3815 \times 23{,}000}{1{,}100 + 0.3815 \times 23{,}000} \times 100 \approx 89$
- $\widehat{CDHES}_{FMLN} = \frac{0.0159 \times 23{,}000}{11{,}000 + 0.0159 \times 23{,}000} \times 100 \approx 3$
- $\widehat{Rescate}_{State} = \frac{0.4729 \times 22{,}000}{1{,}100 + 0.4729 \times 22{,}000} \times 100 \approx 90$
- $\widehat{Rescate}_{FMLN} = \frac{0.0730 \times 22{,}000}{11{,}000 + 0.0730 \times 22{,}000} \times 100 \approx 13$
- $\widehat{TRC}_{State} = \frac{0.4939 \times 7{,}000}{1{,}100 + 0.4939 \times 7{,}000} \times 100 \approx 75$
- $\widehat{TRC}_{FMLN} = \frac{0.0463 \times 7{,}000}{11{,}000 + 0.0463 \times 7{,}000} \times 100 \approx 3$

DWI measures of the relative "dirtiness" of state and FMLN forces vary significantly according to the dataset chosen. State forces clearly committed many more atrocities than FMLN forces in this conflict. But were they thirty times as dirty, seven times as dirty, or twenty-five times as dirty? These radically unstable comparisons suggest the extent to which DWI measures computed from different sources are not, in fact, usefully comparable.

5.3 Timor-Leste: Multiple, Independent Data Sources and Complementary Estimation Methods

In Timor-Leste, Silva and Ball studied conflict-related mortality during the Indonesian occupation using three sources: a convenience sample of almost 8,000 testimonies collected by the official truth commission, a retrospective mortality survey, and a census of public graveyards. [6] [14] They showed that, in such a protracted armed conflict, complementary methods are needed to separate conflict-related mortality into its direct and indirect components.[7] [14]

Silva and Ball found that survey data and methods effectively estimated excess deaths from famine, a population-based phenomenon which accounted for approximately 84,200 deaths. But the population-based survey produced very wide confidence intervals for estimates of direct conflict deaths, an elusive phenomenon. (Using multiple methods, they estimated approximately 18,600

[6] Anonymized data available at `http://www.hrdag.org/resources/timor-leste_data.shtml`.

[7] In this study "indirect deaths" included deaths due to hunger and disease in excess of those which were expected based on the prevailing mortality and population growth rates prior to the onset of conflict. Direct deaths included killings and enforced disappearances.

direct killings and disappearances during the occupation.) In contrast, multiple systems estimation applied to all three available data sources provided reasonably precise estimates of direct deaths, but very imprecise estimates of indirect deaths.

Well-constructed convenience samples may be correlated in broad terms. Conflict-related deaths (including both direct and indirect mortality: killings, enforced disappearances and famine deaths), as documented by the mortality survey and the testimony dataset, are correlated across time at $\rho = 0.97$ and geographic space at $\rho = 0.87$.

But such observed similarities are deceptive: their relative impacts on a DWI would differ by over 350%. Using survey estimates, a DWI measuring the relative magnitude of direct conflict deaths to indirect deaths is approximately $\frac{16{,}090}{102{,}620} \times 100 = 15.7$.[8] The same DWI, constructed using the database of testimonies collected by the official truth commission, yields $\frac{5{,}955}{10{,}809} \times 100 = 55.1$.[9] Put differently, there is more than a three-fold difference in this DWI measure for Timor-Leste based on these two datasets; clearly, the DWI is not robust to measurement differences.[10]

Both convenience and survey methods, and all three datasets (the survey, the testimonies, and the graveyard census) proved necessary to understand the magnitude, pattern and proportional responsibility for the different cause-specific mortality phenomena in Timor-Leste[14]. In Timor-Leste as elsewhere, reliance on individual data sources leads to serious misinterpretations of the relative and absolute scale of conflict mortality. Significant variation in under-registration (by type of violence, region, or perpetrator) results in essentially arbitrary DWI measures for this conflict.

5.4 Colombia's *False Guerrillas*: a cautionary tale

The biases discussed in this paper are generally "side effects" of good-faith data gathering efforts. However, actors contesting dirty wars are often aware of public perceptions of their behavior, and they may therefore attempt to deliberately bias data on violent events. For example, the Colombian Army has been found to create "false guerrillas": noncombatants who are kidnapped, killed, dressed as guerrillas and presented as deaths in combat.[11]

Hicks and Spagat acknowledge that reports of war outcomes can be deliberately biased by parties to the conflict (1662). But note that Colombia's "false guerrillas" directly affect the analysis by perpetrator presented in Hicks' and Spagat's Table 1, "Dirty War Index for Attacks by Actors in the Colombian Conflict, 1988-2005: Civilian Versus Combatant Mortality."

One result of the false guerrilla phenomenon is that the reported number of civilians killed is artificially low, while the reported number of combatants killed

[8] The numerator is the number of direct conflict deaths estimated by the survey $(16,090)$, and the denominator is the number of total direct and indirect deaths estimated by the survey $(16,090 + 86,539 = 102,620)$

[9] This measure is constructed by adding disappearances and civilian killings $(835 + 5,120 = 5,955)$ in the numerator, and these plus excess civilian deaths due to hunger and disease $(835 + 5,120 + 4,854 = 10,809)$ in the denominator.

[10] Neither estimate includes combatant deaths for the Timorese guerrillas and Indonesian Army, for which no reliable estimates exist.

[11] The term used in Colombia is *falsos positivos*, suggesting both positive outcomes for the Army and an incorrect "diagnosis" of guerrilla status. [15]

is artificially high. When this is corrected, the DWI result for the government forces will appear significantly "dirtier" than presented in Table 1. But by how much? As of March 2009, media sources have identified at least one hundred cases; Colombian human rights groups report many more, and have documented this practice in many regions of Colombia over several years.[15] Until the full magnitude of this phenomenon is rigorously estimated, the extent to which Army's data manipulation distorts the DWI remains unknown.

6 Interpretability of the DWI

Hicks and Spagat contend that DWIs "lend themselves to comparisons over time, between wars, between weapons, and between warring combatant groups to identify better versus worse performers" [1: 1658]. Yet indices such as the DWI, in addition to incorrectly measuring rates of prohibited outcomes, may conceal the effects they purport to measure, leading to serious errors of interpretation.

The DWI, like any rate, eliminates absolute magnitudes in order to make situations more easily comparable. This can be a benefit in some circumstances. However, when comparing conflict violence perpetrated by multiple groups, both absolute and relative magnitudes of violence are important measures. Consider a hypothetical case in which one marginal party to a conflict kills ten people during a given period of time, five combatants and five (illegally targeted) civilians. During the same period, a major armed actor has committed 1,000 murders, killing 500 combatants and 500 civilians. The DWI measures these two actors as equally "dirty." In this case, the DWI creates a false—or at least very incomplete—moral equivalence between two very different perpetrators. It is at least equally plausible (though no more measurable) to suggest that the appropriate "moral" measure is a ratio of illegal outcomes to group members.

Hicks and Spagat also claim that over-time monitoring of armed conflict outcomes is a practical advantage of the DWI. However, DWI changes may be extremely misleading. Conflict may "reduce the numerator" by annihilating an at-risk population, such that "dirty" outcomes become decreasingly likely. Furthermore, members of the population at risk may choose to migrate or cease involvement in political activity as a result of earlier violence, decreasing both the numerator and the denominator. Thus, a stable or declining DWI may reflect the success of earlier dirty campaigns, rather than a true reduction in illegal strategies. Over time, the cumulative effects of violence condition the possible levels of "dirty" activities; the DWI is a result of past violence, not a measure of present conditions.

7 Conclusion

Given the urgency of human rights policies and projects, it is easy to understand the impulse to create indices and other metrics, such as the DWI, that purport to evade the ever-present difficulties of collecting and analyzing violence data. Unfortunately, an index created by dividing one manifestly incorrect estimate by another will necessarily be incorrect, and therefore an inappropriate basis for policy-making, analytical comparison or even basic information gathering.

The purported advantages of the Hicks and Spagat DWI measure are based on the false assumption that reporting rates are equal or very similar across different strata of violence, time and locations. If this assumption is not met—and it seldom is—DWI measures reflect the considerable errors and biases of the given single data source, as we showed in examples from the Salvadoran, Guatemalan, Colombian and Timorese cases. We question the basic utility and applicability of DWIs in the context of armed conflict. More generally, we argue that inattention to issues of bias promotes the over-confident, and therefore irresponsible, use of statistics in policy debates affecting human rights and human security.

There are no simple solutions to bad data. Statistical and demographic methods, while complex, still provide the only proven methods for adjusting for and modeling differential coverage of data collection systems, data missingness, inadvertent or deliberate selection bias and, by implication, the true patterns and magnitudes of violence. Yet numbers retain immense rhetorical power and are often immediately necessary. In this context, the solution is investment: in multiple data collection efforts, thoughtful data management protocols, and above all in technologies that can make statistical and demographic methods faster and cheaper.

8 References

[1] Hicks MH-R, Spagat M (2008) "The Dirty War Index: A public health and human rights tool for examining and monitoring armed conflict outcomes." PLoS Med 5(12): e243. doi:10a.1371/journal.

[2] International Committee of the Red Cross (ICRC), Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), 8 June 1977,1125 UNTS 609,, available at: http://www.unhcr.org/refworld/docid/3ae6b37f40.html.

[3] See http://www.hicn.org/research_design/rdn5.pdf.

[4]Guzmán, Daniel, Tamy Guberek, Amelia Hoover, and Patrick Ball. 2007. "Missing People in Casanare" Benetech white paper. Online at http://www.hrdag.org/resources/publications/casanare-missing-report.pdf.

[5] Ball, Patrick, Tamy Guberek, Daniel Guzmán, Amelia Hoover, and Meghan Lynch. 2007. "Assessing Claims of Declining Lethal Violence in Colombia." Benetech. Online at http://www.hrdag.org/resources/publications/CO-PN-CCJ-match-working-paper.pdf.

[6] Silva, Romesh and Patrick Ball. 2007. "The Demography of Conflict-Related Mortality in Timor-Leste (1974-1999): Empirical Quantitative Measurement of Civilian Killings, Disappearances & Famine-Related Deaths." In *Statistical Methods for Human Rights*, J. Asher, D. Banks and F. Scheuren, eds. New York: Springer.

[7] Guberek, Tamy, Daniel Guzmán, Romesh Silva, Kristen Cibelli, Jana Asher, Scott Weikart, Patrick Ball and Wendy M. Grossman. March 2006. "Truth and Myth: Human Rights Violations in Sierra Leone, 1991-2000."A Report by the Benetech Human Rights Data Analysis Group and the American Bar Association's Central European and Eurasian Law Initiative.

[8] Sudman S, Sirken MG, and Cowan CD (1988). Sampling Rare and Elusive Populations. Science, 240(4855): 991-996.

[9] Silva, Romesh and Patrick Ball (2007). "The Demography of Conflict-Related Mortality in Timor-Leste (1974-1999): Empirical Quantitative Measurement of Civilian Killings, Disappearances & Famine-Related Deaths" In *Statistical Methods for Human Rights*, J. Asher, D. Banks and F. Scheuren, eds., Springer (New York).

[10] Spiegel PB and Salama P. War and mortality in Kosovo, 1998-99: an epidemiological testimony (2000). Lancet 355(9222):2204-2209.

[11] Monkkonen, Eric. (2001). "Estimating the accuracy of historic homicide rates: New York City and Los Angeles." *Social Science History* 25(1): 53-66.

[12] *Memoria del Silencio.* The report of the United Nations Commission for Historical Clarification for Guatemala. 1999.

[13] United Nations Truth Commission for El Salvador (1993). From Madness to Hope: The Twelve-Year War in El Salvador. See
`http://www.usip.org/library/tc/doc/reports/el_salvador/tc_es_03151993_casesA.html`.

[14] "Serían 100 los desaparecidosen 9 regiones del país reportados como muertes en combate", *El Tiempo*, 15 October 2008. Online at
`http://www.eltiempo.com/colombia/justicia/2008-10-15/serian-cien-los_desaparecidos-en-nueve-regiones-del-pais-reportados-como-muertos-en-combate_4601931-1`.

[15] Iván Cepeda Castro, "6 de marzo, por las víctimas de los crímenes de Estado," *El Espectador*, 7 March 2009. Online at
`http://www.elespectador.com/opinion/columnistasdelimpreso/ivan-cepeda_castro/columna124967-6-de-marzo-victimas-de-los-crimenes-d`.