Estimating the Human Costs of War: The Sample Survey Approach[*]

By

Professor Michael Spagat
Department of Economics
Royal Holloway, University of London
M.Spagat@rhul.ac.uk

Chapter Prepared for the
Oxford Handbook of the Economics of Peace and Conflict

[*] I would like to thank Hamit Dardagan, Josip Dasovic, Josh Dougherty, Michelle Garfinkel, Madelyn Hicks, Andrew Mack and Paul Spiegel for many useful suggestions. I am solely responsible for the content of this manuscript.

Introduction

In October, 2006 Burnham et al. (2006), hereafter "L2", announced its now-famous sample-survey-based estimate of 601,000 violent deaths in Iraq occurring between March, 2003 and the middle of 2006. Fourteen months later, the Iraq Family Health Study Group (2008a), hereafter "IFHS (2008a)", published results of a different survey that estimated 151,000 violent deaths for virtually the identical time period as that of the L2 survey. The two estimates are fundamentally incompatible with one another; for example, the bottom of L2's 95% confidence interval is nearly twice the top of the IFHS (2008a) 95% confidence interval (426,000 versus 223,000). Fifteen months later, and following a long investigation, the principal researcher of the L2 survey, was censured by the American Association for Public Opinion Research (AAPOR) because he "repeatedly refused to make public essential facts about his research" (AAPOR, 2009a). At the time, Richard Kulka, AAPOR's president, wrote:

> "When researchers draw important conclusions and make public statements and arguments based on survey research data, then subsequently refuse to answer even basic questions about how their research was conducted, this violates the fundamental standards of science, seriously undermines open public debate on critical issues, and undermines the credibility of all survey and public opinion research." (AAPOR, 2009a)

Shortly thereafter, Johns Hopkins University suspended L2's principal researcher from being a principle investigator on any human subjects research for five years (Bloomberg School of Public Health, 2009) after an unidentified body at the University determined that he had violated commitments made to the School's Institutional Review Board to protect the confidentiality of the survey's respondents. There is now a substantial literature dissecting the shortcomings in the L2 survey, with contributions in peer-reviewed journals including Daponte (2007), Johnson et al. (2008), Laaksonen (2008), Rosenblum and van der Laan (2009) and Spagat (2009a and 2010). The authors of the L2 survey have made no substantive response to any of these works.[1]

A series of estimates by the International Rescue Committee (IRC) of "excess deaths" (defined below) in the Democratic Republic of Congo (DRC) rival L2 in their fame, particularly the Coughlan et al. (2006) estimate of 3.9 million excess deaths, subsequently raised to 5.4 million (Coughlan et al., 2008) between August 1998 and April 2007.[2] Yet Human Security Report (2009) closely examines the IRC's published data and concludes that the data themselves can only support an estimate of roughly 860,000 excess deaths with a 95% confidence interval of -550,000 to 2.4 million between May 2001 and April 2007.[3] Moreover, the re-estimate of Human Security Report (2009)

[1] On two occasions inaccurate general responses to critics were posted, and then removed, from the web site of the Bloomberg School of Public Health of Johns Hopkins University (Spagat, 2009a).

[2] However, the journal version of this paper, Coughlan et al. (2009) contains no mention of excess deaths.

[3] The IRC does not give proper confidence intervals for either its 3.9 million estimate or for its 5.4 million estimate, instead offering vague ranges that substantially underplay the uncertainty surrounding their estimates.

treats the IRC data as accurate for the purpose of statistical reanalysis even though the IRC child mortality rate estimates are about twice as high as those of a different and credible survey (Macro International INC., 2009).

Thus, over the last few years two high-profile failures of survey-based estimates of war deaths have shaken confidence in this methodology.[4] Nevertheless, there have been some quieter successes. In the present paper, I reexamine the sample-survey methodology for estimating war deaths and consider the future of the field. Section 2 gives a critical overview of this approach. In section 3 I work through case studies from Kosovo, Darfur, the DRC and Iraq. I draw some conclusions and look to the future in section 4.

The present paper has a number of limitations due largely to space constraints. First, the only human cost I consider is death, leaving out such important costs as injury, displacement and rape. The existing literature focuses strongly on death, perhaps because this is the most dramatic human cost of war.[5] However, the conflict field should allocate more effort into measuring other human costs in the future. Indeed, it is puzzling that there has been so little effort to measure injuries in particular in sample surveys.[6] Injuries could be relatively well measured and they are, arguably, more policy-relevant than deaths since injuries require ongoing treatment and other policy measures. Second, I focus almost exclusively on the survey approach to measuring war deaths since this is an important topic which is already difficult to cover properly within a single chapter. However, a more complete survey of the field would have to include methodologies for measuring war deaths practiced in a number of other projects including those at Uppsala University, International Peace Research Institute, Oslo, Iraq Body Count, B'Tselem and the Benetech Initiative as well as demographic methods.[7]

2. Survey sampling methodology applied to conflict

Theory and sampling

[4] Indeed, survey approaches to measuring deaths due to economic sanctions in Iraq have encountered similar problems. Zaidi (1997) withdrew her survey-based estimate of 567,000 child deaths due to sanctions (Zaidi and Smith, 1995) after revisiting households from the original survey and failing to replicate many of the deaths. According to a subsequent UNICEF survey, child mortality nearly doubled in the early 1990's in Iraq (Ali and Shah, 2000 and Ali, Blacker and Jones, 2003), resulting in an estimated 400,000 to 500,000 "excess deaths" of children. However, Dyson (2009) shows that these estimates are inconsistent with a range of credible evidence and argues that it is likely that the UNICEF "survey data were deliberately manipulated by the then government of Iraq."

[5] Pedersen (2009) gives a broad overview of the literature on health and conflict.

[6] Iraq Living Conditions Survey 2004 (2005b) measured war-related disabilities due to the recent Iraq war. This concept is related to, but not identical with, injuries. For example, some disabilities do not result from violent injuries and some violent injuries do not turn into chronic disabilities.

[7] Some of these approaches are covered in Asher et al. (2008), Brunborg et al. (2006), Hicks et al. (2008) and Uppsala Conflict Data Program (2010).

We could try to estimate the total number of people who have died in a conflict during a particular time period by carrying out a census of the entire conflict-affected population. This approach requires conducting interviews at every household within the conflict-affected area in an attempt to record the fate of every single person. However, conflict censuses are extremely expensive and will, consequently, always be rare.[8]

An obvious money-saving alternative to a census is a sample survey which, in theory, should be able to underpin good estimates of conflict deaths as long as samples are sufficiently large and representative of a full conflict-affected population. Samples may provide some further advantages relative to censuses. For example, the smaller scope of a survey could make it possible to focus more on data quality than a census might be able to do, although in the final analysis this comparison would depend on how well funded the two approaches are relative to their scope.

The idea behind the use of sampling to estimate war deaths is straightforward but worth a brief discussion in the interest of clarity. Suppose, for example, that we draw a sample such that every household in the affected population has an equal probability of selection into the sample.[9] Suppose, further, that for each household in our sample we manage to measure accurately both the number of living household members and the number of former household members who have died due to a war.[10] If x% of the sample population are found to have died due to the war then x becomes an unbiased estimate for the percent of war deaths in the full war-affected population. If we know the size of this population then we can multiply by this number to obtain an unbiased estimate of the total number of war deaths in this population. With additional details about the sample design we can also construct a confidence interval that quantifies the sampling (but not the non-sampling) error around this estimate. Unfortunately, these steps are fraught with potential pitfalls even though they are simple and few.

The quality of any survey depends crucially on how its sample is built yet it is often very challenging to draw a good sample. Researchers will be lucky to have a reasonably complete list of the households comprising the population affected by a conflict. Such a list may exist if, for example, there has been a recent census. If so, then it will be possible to draw a *simple random sample* which can be conceived as follows.

---

[8] Moreover, even a national census might fail to measure conflict mortality adequately in cases where victims, and possibly perpetrators, of violence have been driven abroad. In such cases good measurement requires access to these displaced populations. Many attempts to measure conflict mortality will need to contend with such scattering of the affected population but the costs of administering a census are particularly sensitive to this problem.

[9] It is not necessary that every household have equal probability of being chosen. We can construct valid estimates as long as each household has a known and non-zero probability of being chosen.

[10] Here we ignore all problems in defining households, although doing this is particularly tricky and problematic in a conflict environment. There are births, deaths, in-migration, out-migration, cases of multiple families sharing single segmented dwellings, household mergers and splits, forced migration and other phenomena that complicate definitions of households.

Assign a number to every household in the affected population. Write each number on otherwise identical balls and place them in a huge urn. Draw balls at random and do interviews with the households that correspond to these balls.

A big problem with simple random sampling, even when it is feasible, is that it can be prohibitively expensive to conduct interviews at all of the households selected with this method. The conflict-affected population may be dispersed over a large territory. In this case, interviewing all households in a simple random sample might require field teams to travel great distances, perhaps hundreds of miles over rough terrain just to perform one interview. A common cost-cutting response to this logistical problem is to interview *clusters* of nearby households. Concentrating interviews close together reduces travel time and other expenses.[11] This is the main practical reason why cluster sampling is widely used in conflict mortality survey research.

The use of cluster sampling rather than simple random sampling will normally widen confidence intervals because households located close to one another are likely to have similar conflict mortality experiences. This effect can be quite large in small cluster surveys since a few unrepresentative clusters in a survey with a small number of clusters can give a very misleading estimate (Spagat 2009b). Nevertheless, comparing a cluster sample to a simple random sample with the same number of households can itself be misleading, since it will not in general be feasible to switch from the former to the latter on a fixed budget. From the economic point of view, cluster sampling, by enabling a larger sample size on a fixed budget, may well outperform simple random sampling.

A serious issue remains over how to select a sample when there is not a reasonably reliable enumeration of households in the affected population. A variety of techniques have been used to tackle this problem. Briefly, the key step in these methods is to somehow carve up a geographical space into manageable units such as towns, villages or city neighborhoods. Next, population estimates for these units are used to select some of them with probability proportional to these population estimates. Of course, sample quality will depend very much on the quality of these population estimates. At this stage at least two broad lines are followed. The first, more careful, technique is to fully enumerate households within the chosen units. Although making such listings takes time, this procedure is still considerably cheaper than enumerating all households within the full affected population, i.e., essentially conducting a census. Households are then selected randomly for interviews from these specially-created within-cluster household lists, often with the use of simple random sampling. The second class of approaches, considerably cheaper than the first but far less accurate, usually involves some kind of directional sampling. For example, teams begin from some central point in the geographical space, spin a pen, select a first household at random along the direction the pen points and continue conducting further interviews at other households that are near to the first one (Grais et al. 2007).[12] A serious problem with such

[11] A second potential advantage of cluster samples relative to simple random samples is that by reducing travel time, cluster samples reduce exposure of interview teams to risk. This can be important since conflict surveys are often conducted in relatively dangerous environments in which travel entails risk.

[12] Johnson et al. (2008) argues that the directional sampling procedure used in Burnham et al. (2006) may have introduced a substantial upward bias.

procedures is that they do not normally determine households to be interviewed unambiguously, thus allowing subjective judgments of interview teams to express themselves. For example, in an urban environment it will often be impossible to travel very far in a direction selected by a pen spin so it will be necessary to change course and it is unclear how this should be done. Proximity is rarely, if ever, defined in write-ups of surveys, although various definitions are possible. Lack of clear guidance can allow field teams to use their own discretion in moving from house to house, undermining the random selection of households. In general, little is known about the properties of directional sampling, making these a very important topic for future research.

Another crucial issue is the relationship between the population group actually covered by surveys and to the full conflict-affected population that is of interest to researchers and the general public. Some conflict mortality surveys are conducted exclusively, or largely, within special settlements of displaced people. Sometimes, particularly when surveys are conducted during a conflict, some parts of affected populations cannot be safely accessed. In such cases the population actually covered by a survey is only a sub-group of the full conflict-affected population that we are interested in and we cannot know how representative the former is of the latter. Populations displaced by a conflict will, almost by definition, have suffered some conflict exposure, thus suggesting that their mortality rates might be above-average relative to the full conflict-affected population. On the other hand, displaced people have managed to reach a camp, possibly because they have access to greater-than average resources or faced lower-than-average risks compared to people who died. Thus, presence in a camp may actually be correlated with lower-than-average mortality rates. In fact, there are so many potential biases present in groups of displaced people that there is generally no way to know how representative a displaced population might be of a broader conflict-affected population. This means that extrapolations from sub-samples of displaced people are dangerous and should not be attempted.[13] Similarly, while it may be tempting to assume that areas that are currently inaccessible to survey-research teams have experienced more violence in the past than currently-accessible areas, this assumption might not actually be true. For example, an armed group may have established dominance in an area in the recent past and currently be able to offer secure access for researchers. But the area might have experienced very high mortality while this armed group was establishing its control in the first place (Kalyvas, 2006). The reverse can also be true. An area that is presently contested, and thus inaccessible to survey teams, might have been relatively safe in the recent past. Once again, projecting results from surveyed areas onto areas that have not been surveyed can be misleading.

What is measured

Three broad classes of war-death-related estimates have been published in the literature: violent deaths, mortality rates and excess deaths (which are based on mortality rates). Violent deaths are usually assumed to be war-related, although some violent

[13] General Accounting Office (2006) examines widely discrepant estimates of war deaths in Darfur and argues that many of these divergences probably result from improper extrapolations that have been made from displaced populations onto general populations.

deaths probably would have occurred even without war. Sometimes estimates are presented simply of mortality rates. These are expressed in a variety of units such as deaths per 10,000 per day or deaths per 1,000 per year. The excess-deaths concept is meant to measure deaths, both violent and non-violent, that would not have occurred if a conflict had not occurred. This means that the excess-deaths concept is, fundamentally, based on a counterfactual analysis, making all excess-death estimates rather speculative in nature. The key to any excess-death estimate is to establish a plausible baseline mortality rate to serve as a counterfactual. In practice, these have been taken to be either a regional average or else a pre-war mortality rate has been projected forward as an assumed rate that would have prevailed in the absence of war. Neither assumption is very plausible. Many countries will tend to be systematically either above or below average for their regions independent of whether or not they suffer a war. Moreover, there is no reason to assume, as is implicit in the excess-deaths concept that all changes in mortality rates that coincide with a war are specifically caused by the war. For example, a drought may cause a war while directly causing much of an increase in mortality accompanying the drought/war. In this case it would be highly misleading to treat the war as the sole cause of the increase in mortality as typical excess-deaths calculations would do. Consequently, Human Security Report (2009) argues that the foundations of excess-death calculations are so weak that they should be abandoned, except possibly in a handful of cases for which data are particularly rich.

Field work[14]

An evaluation of the details of how survey data are gathered in the field is the other key ingredient for understanding the quality of a survey. This dictum applies in particular to conflict surveys, since they are frequently conducted within highly charged political environments and are inextricably intertwined with the assignment of blame for violence and atrocities. Members of a particular religious, ethnic, social or class group may wish to pin blame for killings on a rival group and such motivations can affect behavior of both interviewers and interviewees alike. At one extreme either interviewers or interviewees might simply invent deaths that have not occurred.[15] Or blame for real deaths may be shifted by either party from a group that was actually responsible, or from an unknown group, to a different one. Some surveys reduce the above motivation for information distortion by not asking interviewees to affix group blame for violent deaths.[16]

Both interviewees and interviewers may treat a survey as a tool to attract international aid or to encourage the intervention or withdrawal of international powers from the conflict zone (General Accounting Office, 2006, p. 15). Depending on the situation, such motivations could lead to either under or over-estimation of the number

[14] Asher (2009) gives a good treatment of field work issues in conflict surveys.
[15] Spagat (2010) presents evidence of fabricated deaths in the Burnham et al. (2006) survey.
[16] For example, the questions recommended by SMART Methodology Version 1 (2006) for determining causes of war-related deaths will not determine the group membership of perpetrators (p. 78).

dead and to other distortions in survey findings. Interviewees or interviewers may perceive that a large estimate of deaths is necessary to claim the attention of a jaded and distracted international community, to focus opprobrium on parties to a conflict or to attract aid or reparations. Alternatively, interviewees or interviewers may think that low mortality numbers might help to avoid external intervention or to maintain an ongoing intervention that could be viewed internationally as having a positive effect if current deaths rates are low. Indeed, the write-ups of many conflict mortality studies contain pleas for interventions or criticism for the absence of past intervention. Such advocacy may be appropriate but, simultaneously, highlights the need for careful scrutiny of the quality of data collection efforts and how these might relate to outcomes desired by survey participants.

More mundane concerns, common to all survey research, also apply with particular strength to conflict surveys. Interviewees might not follow proper procedures for a variety of reasons. One extreme behavior is to simply make up answers to survey questionnaires without actually doing the interviews (AAPOR and ASA, 2003). Such cheating can sometimes be detected by inspection of completed questionnaires and computer programs can also be used to detect patterns suggestive of fabrication (Bredl, Winker and Kötschau, 2008). Interviewers may also cheat in more subtle ways, for example, doing interviews in a more convenient location than a remote location selected by sampling procedures. Or interviewers may simply get careless at the end of a long day if they are under pressure to do many interviews very quickly. The motivation to cut corners can be especially strong in conflict surveys where field work is likely to be physically strenuous and possibly dangerous. Thus, it is particularly important in conflict research to implement a solid system of quality checks. These can include comparing results turned in by different interviewers and field teams with each other, contact of a sample of households who are supposed to have been interviewed to make sure they were actually interviewed and re-interviews of a random sample of households by a second team to check on the quality of the first team's work (AAPOR and ASA, 2003). It is important that field teams know in advance that they will be subjected to such scrutiny so that they are on their best behavior from the very beginning.

The backgrounds of interview teams and how these relate to those of the interviewees also need to be considered. If interviewers are readily linked to one side in a conflict then interviewees might tune their responses accordingly. Another issue is that local culture might preclude candid interviews of males by females and vice versa.

Interviewees might not be politically motivated but may struggle to remember information accurately or may under report unpleasant information. For example, some analysts think that that interviewees tend to underreport deaths of small children (Sullivan et al., 2000) although Hill and Choi (2006) find little evidence to support this view. People might remember real deaths but shift the timing of these deaths unintentionally. Timing inaccuracies can be important for a survey if, for example, pre-war deaths are reported as post-war deaths inflating the death count attributed to a war. Since memory fades with time, surveys covering long time periods will be less accurate than comparable surveys covering short time periods. Household composition may also be rendered increasingly inaccurately as recall periods are extended. Extended family members might circulate through a household at various points in time, especially in families that have

experienced displacement. Mortality rates could be exaggerated if deaths of temporary household members are treated as deaths of permanent household members.

3. Surveys of particular conflicts

Kosovo

Spiegel and Salama (2000) used a cluster survey with 50 clusters and 24 households per cluster to estimate 12,000 deaths due to "war-related trauma" with a 95% confidence interval of 5,500 to 18,300 for the war in Kosovo between February, 1998 and June, 1999. Researchers worked from a 1991 census, adjusted using updating information from various sources, to randomly select 50 villages or neighborhoods. Although they were not able to include villages with populations below 100 the survey does seem, nevertheless, to have managed rather good coverage of the affected population. Field teams started from the center of each unit, selected a random direction by an unspecified method and proceeded to the edge of the unit, recording either the distance to the edge or the number of housing units along the way. They then chose a random household along this radius for their first interview in the cluster and moved to the right from nearest household to nearest household until they completed 24 interviews. There are some ambiguities in the description of the sampling procedures. The description seems to assume discrete houses so it is unclear how apartment buildings were handled. Also it is unclear how the center of each unit was defined and exactly what it means to select a random direction from the center. For example, in an urban environment the street layout may allow only two or four feasible directions of movement out of the center. There could be many areas unreachable to the survey depending on how this ambiguity is resolved. If mortality in the unreachable areas differs significantly form mortality in the reachable areas then the sample could be biased. For example, it is possible that ethic cleansing campaigns perpetrated by the Yugoslav government penetrated less deeply into relatively inaccessible areas than into accessible ones. On the other hand, it seems much less likely the NATO aerial bombardment campaign would have followed such a pattern.

The authors give little information on field work other than that it was done by 14 field teams each consisting of 2 Albanian speakers. It is possible that some Albanian interviewees or Albanian-speaking interviewers may have wanted to exaggerate the number of deaths to call attention to the victimization of their group. Such incentives would have been weakened, but not eliminated, by the fact that interviewees were not asked to identify perpetrators of violent deaths, thus denying them a direct opportunity to blame Serbs for their relatives' deaths. The authors do not describe any quality-control procedures such as comparing interview results across teams, checks by supervisors that interviews were done or re-interviews to make sure interviews were done properly. The recall period was 17 months which seems to be quite reasonable.[17]

[17] SMART Methodology Version I (2006) states categorically that recall periods longer than a year should not be used, but this would appear to be an inordinately conservative rule.

A notable and unexpected finding of the study is extremely high war-related death rates for males above 50 years old, who are estimated to have been killed at more than three times the rate for military-aged males. If true, this suggests that older men may have been particularly targeted in the ethnic cleansing campaign.

Darfur

Depoortere et al. (2004) did cluster surveys at four sites within West Darfur that included internally displaced people (IDPs). The surveys measured mortality of households both at the sites and, for three of the four, before arrival to the sites. Some people at these sites were living within special camps for the displaced while in other areas displaced people were mixed in with the permanent population. Researchers did 30 clusters of 30 households at three sites while in the fourth, with limited time, they did 30 clusters of 15 households. The authors follow good practice in not extrapolating their results to the full affected population of West Darfur or all of Darfur. Their sub-population of IDPs is of valid interest by itself.

Sampling procedures were similar to those in Spiegel and Salama (2000). Relatively small slices of territory were chosen. In each one a field team walked from the center to the edge in a random direction selected by an unspecified method, enumerating all the households in between. They then selected one household at random and did interviews at that one and 29 (or 14) further proximate households. An apparent weakness is that "proximity" is left undefined in the paper, opening the possibility that interview teams may have had discretion in their house-to-house movements. Well-specified sampling procedures rule out such discretion so that interviewers are not free to gravitate, consciously or unconsciously, in directions that might look like they will yield results the interviewers would favor.

The description of the sampling methods does not strongly suggest that there were unreachable parts of surveyed areas. It is, however, likely that locations close to cluster centers had higher selection chances than areas at the edges.[18] Imagine a field team setting out from the center of a roughly circular camp. There is a much narrower range of random directions that will lead the team to pass, and hence list, a household located near the edge of the camp than is required for the team to pass and list a household near to the center. Thus, this sampling scheme can introduce bias if mortality rates near camp centers differ systematically from rates at the edges. Suppose, for example, that camps are formed by successive waves of displaced people where the first wave defines the center and later waves extend the camp outwards more or less in concentric circles. Mortality may differ systematically between these waves, although there is no reason to expect that the mortality rates of later waves should be either higher or lower than earlier ones. Indeed, an interesting finding of Depoortere et al. (2004) is that mortality experiences of neighboring families within camps, *even pre-arrival*, tend to be very similar. This suggests that spatial patterns of mortality with IDP camps could be quite complex and subtle, implying that the details of sampling schemes can be crucial in determining results.

[18] SMART Methodology Version 1 (2006) makes this point on page 56.

An undisclosed number of interviewers, described as "local, highly literate", accompanied by Arabic-speaking translators of unspecified backgrounds conducted the interviews. It is unclear how these interview teams would relate to the interviewees. As the interviewees were current and prospective recipients of international aid they might have perceived that they had an interest in exaggerating their mortality experiences and interviewers might have shared the goal of attracting international action. The main quality check on the field work appears to be that filled-out questionnaires were checked for accuracy each day. Recall periods were very short and conservative, ranging between 39 and 193 days.

During the pre-arrival periods the central estimates for the crude mortality rates were in a range of 5.9 to 9.5 per 10,000 people per day with upper and lower limits for 95% confidence intervals between 2.2 and 15.7 per 10,000 per day. All these numbers far exceed a common emergency threshold of 1 per 10,000 per day. Even within-camp crude mortality rates tend to exceed this threshold, in one site by at least a factor of four. In short, Depoortere et al. (2004) left little doubt that there was a humanitarian crisis in Darfur. A further result is that although most people violently killed were estimated to be adult males, significant numbers of women and children were also killed.

There is great public interest in numbers for violent or excess deaths for the Darfur conflict. However, there are no surveys with proper random coverage of all of Darfur at any time during the conflict and no obvious mortality rate to use as a baseline in an excess-deaths calculation. Nevertheless, a supply of Darfur estimates has arisen to satisfy the demand for a single number to sum up the human cost of the war. GAO (2006) evaluates six such disparate attempts that cover different time periods and propose figures ranging from 63,000 excess deaths to 400,000 total deaths. Hagan and Palloni (2006) and Degomme and Guha-Sapir (2010) made subsequent estimates, still in the absence of a proper random sample or a clear baseline. Despite these problems Degomme and Sapir (2010) do usefully bring to bear evidence from 63 local surveys and shows rather convincingly that mortality rates have dropped strongly between 2004 and 2008 with violence rates decreasing much more quickly than diarrhea-related mortality.

Democratic Republic of Congo (DRC)

Beginning in 2000 the International Rescue Committee (IRC) has released a series of five estimates of excess deaths in the DRC covering longer and longer periods: 1.7, 2.5, 3.3, 3.9 and 5.4 million excess deaths have been estimated with the last figure covering August 1998 to April 2007.[19] Human Security Report (2009), hereafter "HSR", examines these estimates in great detail and comes to two main conclusions. First, the IRC's central estimates are too high by a very wide margin. Second, carefully calculated confidence intervals are so wide that the IRC's central estimates are not very meaningful. In this section I mainly summarize the analysis of HSR and then take a closer look at Coghlan et al. (2006) since this work illustrates in microcosm some of the main problems with the IRC estimates and was the only study in the series that was published in a well-known peer-reviewed journal.

[19] These numbers come from, respectively, Roberts (2000), Roberts et al. (2001), Roberts et al. (2003), Coghlan et al. (2006) and Coghlan et al. (2008).

Roberts (2000) and Roberts et al. (2001), the first two studies in the IRC series, produce estimates based on eleven separate sample surveys of eight areas of the DRC that are non-randomly selected and very small. Because of the non-random selection mechanism these surveys are not appropriate for extrapolation to estimates of excess deaths for all of the eastern DRC. Yet the IRC does this anyway with its estimates of 1.7 and 2.5 million excess deaths.[20] HSR argues, correctly, that these estimates should be disregarded and that the IRC data are not usable for region-wide estimates for the period August 1998 to March 2001.[21]

For the period between May 2001 and April 2007 the IRC surveys do have reasonable national coverage and are, therefore, usable to estimate national mortality rates. The IRC pushes one step further by estimating excess mortality. As noted above, this requires specifying a hypothetical baseline mortality rate and assuming that the DRC would have experienced this rate if there had not been a war. The IRC assumes, implausibly, that without war the DRC would have experienced the average mortality rate for all of Sub-Saharan Africa: 1.5 per 1,000 per month.[22] This is not credible because decades of misrule by the government of Mobutu Sese Seko, rendered it very unlikely that the DRC could have achieved the mainstream mortality rate for all of Sub-Saharan Africa. Of course, one cannot know an appropriate counterfactual mortality rate. This is the main reason why the excess-deaths concept is problematic in the first place. However, for the sake of argument, HSR proposes a baseline of 2.0 per 1,000 per month, which is what the IRC itself measures in western DRC. This region was largely untouched by the war and is, therefore, a plausible candidate to represent what the east might have experienced without war. The mortality rates measured in the three surveys with national coverage conducted by the IRC turn out to be rather close to this alternative baseline. Consequently making this reasonable change to the baseline mortality rate reduces the excess-death estimate from 2.8 million to 900,000, i.e., by more than a factor of three.

The IRC has never published a proper confidence interval for its excess-death estimates.[23] It does, however, give confidence intervals on its mortality-rate estimates which make it possible to construct confidence intervals for excess-death estimates. HSR does these calculations and finds that, if you accept the IRC's proposed baseline as an absolute certainty, the 95% confidence interval on excess deaths becomes 1.3 million to 4.5 million. If, instead, you accept the HSR's proposed baseline then the 95% confidence interval becomes -550,000 to 2.4 million, meaning that the IRC data is not even robust

---

[20] Note that the IRC does exactly what Depoortere et al. (2004) properly refrains from doing: extrapolate from surveys of non-randomly chosen areas to a region-wide estimate.
[21] Human Security Report (2009) also shows that even if one does assume, contrary to the reality, that the five areas in Roberts (2000) were randomly chosen a correct estimate would still be only about half of the IRC one. The Roberts (2000) estimate is inflated by giving disproportionate weight to a tiny area with an exceptionally high death rate.
[22] The IRC lowers this baseline several years later to 1.4 per 1,000 per month in Coghlan et al. (2008).
[23] The IRC does offer some restricted ranges based on running a few scenarios under varying assumptions but never shows real confidence intervals calculated in a standard way.

enough to reject a hypothesis of negative excess deaths at a standard significance level. Assuming a probability distribution running between the IRC and HSR baselines would, of course, produce an even wider confidence interval than either of the above ones. Moreover, at best these confidence intervals incorporate only sampling error while non-sampling errors are likely to be large as well.[24] In short, the IRC estimates are so imprecise that they have little meaning.

The problems with DRC excess mortality baselines and confidences intervals are clearly illustrated in microcosm by Coghlan et al. (2006). The estimate for the national crude mortality rate during the period January 2003 through April 2004 is 2.1 per 1000 per month with a 95% confidence interval of 1.6 to 2.6. Under the IRC's baseline assumption this calculation gives an excess death rate estimate of (2.1 – 1.5) = 0.6 per 1000 per month for a 15-month period. Applying this rate to the IRC's population estimate of 63.7 million gives (0.6 excess deaths per 1,000 per month) x (15 months) x (63,700,000 people)/(1,000) = 573,300 excess deaths which rounds up to 600,000.[25] Continuing to treat the baseline as completely certain but incorporating the IRC's 95% confidence interval over the mortality rate of 1.6-2.6, the 95% confidence interval on this estimate becomes 100,000 to 1.1 million. If we replace the IRC baseline with the HSR baseline then the central estimate becomes 100,000 with a 95% confidence interval of -400,000 to 600.000. Thus, changing the baseline rate in a plausible way reduces the IRC's central estimate by a factor of 6. Moreover, note that the ranges on these confidence intervals are very large compared to the central estimates themselves and that it is not possible to rule out that excess deaths have been negative.

Note further, that all the above analyses accept the IRC mortality rate estimates as given. However, these may well be too high. The IRC child mortality estimates are, in fact, roughly twice as high as those measured by a Demographic and Health (DHS) survey (Macro International Inc., 2007) which are generally regarded to be very high-quality surveys (Human Security Report, 2009). Both of these estimates cannot be correct.

Iraq

There have been five prominent survey estimates of violent deaths or excess deaths in the ongoing Iraq conflict. These come from Roberts et al. (2004), hereafter "L1", the Iraq Living Conditions Survey (ILCS, 2005), Burnham et al. (2006) (L2), IFHS (2008a) and the estimate of the polling firm Opinion Research Business (ORB, 2008). As noted above, the violent-death estimates of L2 and the IFHS fundamentally conflict with one another, with the bottom of L2's 95% confidence interval nearly twice the top of the IFHS 95% confidence interval. This is a simple comparison because the L2 and IFHS surveys cover virtually the same time periods. Comparisons are more difficult for surveys covering different time periods because violent-death rates vary substantially

[24] Important non-sampling errors include uncertainty over the baseline mortality rate itself, inaccuracies in establishing household boundaries and misunderstood questions perhaps due to translation issues.

[25] This calculation ignores, as the IRC does, the fact that 8% of the population was inaccessible to the survey.

over time. Therefore, to facilitate such comparisons we incorporate the other widely quoted source on violent deaths in the Iraq war: Iraq Body Count (IBC). IBC records civilian deaths based on monitoring a large number of sources including the media, hospitals, morgues, NGO's and governments (Hicks et al., 2008). IBC data are readily comparable to those from any of the five surveys because IBC data is daily, covers the entire conflict and is compiled using a uniform methodology.

Table 1, taken from Spagat and Dougherty (2010), compares the five surveys along various dimensions, with IBC units serving as a measuring rod in some of the rows. The huge discrepancies in violent-death estimates apparent in the table has to call into question the validity of the survey approach for measuring violent conflict deaths. At least some of these surveys have to be wrong by wide margins. To maintain confidence in the survey approach to estimating violent conflict deaths we need to explain which of these surveys are wrong and why this is the case. Otherwise, it will be impossible to avoid similar errors in the future. Below I sort through the surveys in search for a viable path forward.

Table 1. Five Surveys of Violent Deaths in Iraq since March 2003

	ILCS	L1	IFHS*	L2	ORB
Coverage Period Ends	May 1, 2004	September 20, 2004	June 30, 2006	July 10, 2006	August 31, 2007
Survey Estimates of Violent Deaths	26,000	56,700	98,000 or 151,000	601,000	1,033,000
Ratio to IBC	1.7	3.0	2.0 or 3.1	12.2	12.2
Violent Deaths in Baghdad	8,063	18,900 – 27,000	52,920 or 81,540	150,000	600,000
Ratio to IBC in Baghdad	1.0	1.9 – 2.7	1.9 or 2.9	5.2	12.5
Number of clusters	2,200	32	971	47	112
Field Questionnaire Available	Yes	No	Yes	No	No
Household Roster Taken	Yes	Yes	Yes	No	No

Source: Spagat and Dougherty (2010)

*The IFHS argued that there is a general tendency for the underreporting of deaths in household surveys. On these grounds, the IFHS adjusted its estimate upwards by more than 50%, whereas all the other surveys in this table used conventional estimatation methods without such an adjustment. It is, therefore, best to remove this adjustment when comparing across surveys so in the IFHS column I always provide two estimates; the first is a conventional estimate and the second is an adjusted one as given in the IFHS.

Based on national estimates, the five surveys separate naturally into two groups. On the one hand, there are the ILCS, L1 and the IFHS with estimates roughly two to three times IBC figures for comparable time periods.[26] On the other hand, the L2 and ORB estimates both exceed IBC figures by more than a factor of 12. Focusing only on Baghdad the L2-ORB grouping largely breaks down while the ILCS-L1-IFHS grouping holds up fairly well. The bottom three rows of Table 1 highlight a few quality indicators suggesting that the ILCS-L1-IFHS estimates are likely to be much closer to the truth than are the L2-ORB ones. L1, L2 and, to a lesser extent ORB, all have a small number of clusters and are, therefore, vulnerable to drawing unrepresentative samples that include too many high-violence clusters (Spagat, 2009b). The ILCS and IFHS are much larger surveys that do not have this weakness. Moreover, the ILCS and IFHS are the most open about their methodologies. This is exemplified in Table 1 by the simple fact that, alone, the ILCS and the IHFS meet the extremely minimal standard of disclosing their questionnaires. In recent years there have been attempts within the survey profession to establish standards for disclosure of essential methodological information (AAPOR, 2006). In fact, the lead author of the L2 survey, was found to be in such serious breech of these standards that he was formally censured by the American Association for Public Opinion Research, one of only three formal censures applied over a 12-year period.[27] ORB also refuses to disclose its questions as asked in the field, although ORB has at least released an English version of these questions. Finally, the table focuses on one aspect of methodological weakness of L2 and ORB relative to the other surveys: a failure to compile a list of all members of each household in the sample (a household roster) together with some basic demographic information on each household member.[28] Not taking household rosters is generally viewed as bad practice (e.g., SMART Methodology, 2006, p. 75).

Further analysis of the L2 and ORB estimates reinforces the view that these are out of line with reality. Spagat and Dougherty (2010) focus on the ORB poll. A key finding is that collation of ORB data ranging across three separate polls, including the one underpinning ORB's violent-death estimate, reveals critical inconsistencies suggestive of compromised data collection. More than 80% of ORB's estimated deaths come from just four contiguous governorates of Iraq. Within these governorates a greater percentage of respondents report deaths of *household* members than report deaths of *extended-family* members in a separate ORB poll taken only six months earlier. This is not a credible pattern since extended-family networks reach far beyond household boundaries. Respondents in the southern governorates of Iraq do, in fact, display the expected pattern of far fewer deaths of household members than deaths of extended-family members. But, incredibly, the center moves in the opposite direction and accounts for the vast majority of ORB's estimate. Spagat and Dougherty (2010) argue further that there are many key quality shortcomings in the ORB poll such as ambiguous questions

[26] Part of the differences between IBC and the surveys comes from the fact that IBC counts only civilian deaths whereas the surveys include both civilians and combatants.
[27] AAPOR (2009b) lists the methodological details that the principal researcher on L2 has refused to disclose.
[28] Spagat (2010) includes evaluations of the methodological quality of the ILCS, IHFS and L2 and Spagat and Dougherty (2010) evaluates the quality of the ORB survey.

that are inadequate to prevent respondents from reporting non-violent deaths or deaths of extended family members, an unsound treatment of non-response and incorrect calculation of confidence intervals. In short, the evidence suggests that the ORB estimate is very unreliable and far too high.

The departure of the L2 survey from the broad body of evidence on violent deaths in Iraq goes well beyond the comparisons presented in Table 1. Here I highlight just two illustrative examples on the temporal and spatial patterns of violent deaths.[29] For the period June 2005-June 2006 the violent-death rate measured by L2 exceeds that of the IFHS by more than a factor of seven with the bottom of L2's 95% confidence interval nearly triple the top of the IFHS one (IFHS, 2008a). The L2 estimate in five central governorates of Iraq exceeds the ILCS one by a factor of nearly twelve and even exceeds the top of the ILCS 95% confidence interval by a factor of 7.5 (Spagat, 2010).

There have been various attempts in the literature to explain the errors in the L2 survey. Rosenblum and van der Laan (2009) note that violence levels vary strongly across clusters in the L2 sample and that the sample size is small. These factors imply that the sample distribution of violence by cluster could diverge strongly from the underlying distribution from which the data are drawn. They argue that under these circumstances it is inappropriate to calculate confidence intervals either by assuming a normal distribution or through Monte Carlo methods. They propose an alternative method for calculating confidence intervals in such environments that allows for divergence between underlying patterns generating the data and patterns apparent only in sample data. They apply this technique to recalculations of confidence intervals for the L2 data, all of which come out much wider than the published confidence interval of Burnham et al. (2006); one even extends below 100,000. In a similar spirit, Spagat (2009b) performs simulations using ILCS data, suggesting that small cluster surveys like the L2 one are unreliable and can easily overestimate violent deaths by a factor of three or more if they select a few unusually violent clusters. Johnson et al. (2008) and Onnela et al. (2009) argue that the final-stage sampling methods used in L2 are biased towards violent areas and that this bias could cause great overestimation, even by a factor of three for plausible parameter values in the model. Spagat (2010) provides evidence of data fabrication and falsification in the L2 survey that could account for the strong discrepancies between the L2 estimate and other evidence.[30]

Thus, close examination of both the ORB and L2 surveys provides ample reason to discard both of these estimates. The remaining three surveys are broadly consistent with one another (Table 1). However, the L1 estimate is extremely imprecise due to its very small number of clusters and, hence, this survey should receive little weight.[31] Thus, we are left, essentially, with the ILCS and the IFHS.

Available documentation (ILCS, 2005, IFHS, 2008a, IFHS, 2008b) suggests that the data were carefully gathered for both the ILCS and IFHS surveys. The quality of the ILCS field work benefited from the fact that its field work was carried out during a

[29] More details can be found, e.g., in Spagat (2009a) and Spagat (2010).
[30] See the conclusion of Spagat (2010) for a concise summary of this evidence.
[31] Roberts et al. (2004) gave a confidence interval of 8,000 to 192,000 excess deaths for the country excluding one governorate and does not give a confidence interval for any estimate with this governorate included.

relatively peaceful time in Iraq. On the other hand, a weakness of the ILCS for war-death estimation is that the follow-up cause-of-death question for respondents reporting deaths is not ideal, because it forces a choice between "disease", "traffic accident", "war-related", "during pregnancy, childbirth or within 40 days after", "other (specify)" or "don't know". There would appear to be a close correspondence between war-related deaths and violent deaths but it would be better if these were pinned down more clearly in the ILCS questionnaire.

A clear problem with the IFHS is that, due to security reasons, the field teams failed to visit 115 out of 1086 selected clusters. In particular, the teams visited just 37 clusters in 108 attempts in Anbar, 65 out of 96 in Baghdad, 60 out of 72 in Nineveh and 53 out of 54 in Wassit. IFHS (2008a) adjusts for missed cluster visits only in Baghdad and Anbar and gives no reason why it did not attempt an adjustment in Nineveh. For the Baghdad adjustment IFHS (2008a) assumes that the ratio of IFHS figures to IBC ones in Baghdad should be the same as the IFHS/IBC ratio in six high-mortality governorates. IFHS (2008a) implements this assumption by imputing to the missing Baghdad clusters the violence levels necessary to bring the IFHS/IBC ratio in Baghdad into equality with the IFHS/IBC ratio in the six high-mortality governorates.[32] The implications of this assumption turn out to be strong, since equating these two ratios requires assuming that the missed clusters in Baghdad are four times as violent as the visited ones over the entire period of 3.3 years covered by the IFHS estimate. This is a very strong assumption since it is based only on an observation (inability to visit a cluster) taken at a single point in time. The impact is to add 30,000 violent deaths to the final IFHS estimate beyond what it would be if the missing clusters were equally violent as the visited ones. Implicit in this adjustment is an assumption that the percent of violence covered by IBC is the same in Baghdad as it is in the six high-mortality governorates. Thus, this upward adjustment would be too large to the extent that IBC's coverage in Baghdad is actually better than its coverage elsewhere.[33] When the IFHS dataset is released it will become possible to reevaluate this missing-cluster adjustment using district-level IBC information together with the breakdown of missing clusters by district in the IFHS. Pending this reanalysis, the IFHS Baghdad adjustment should be regarded with some skepticism.

Another issue concerning the IFHS estimate is that it is the only survey-based estimate of conflict deaths in the literature that has been adjusted upwards to account for presumed underreporting of deaths. This adjustment is large, about 54%, and causes two problems. First, this unprecedented adjustment complicates comparisons with other surveys since IFHS (2008a) is the only survey that makes such an adjustment. The case that IFHS (2008a) makes for its adjustment is generic to all conflict surveys. Therefore, if it is correct to adjust the IFHS upwards for these reasons then it is also correct to make

[32] The same procedure is used for the Anbar adjustment but, unlike in Baghad, the violence levels imputed to the missing Anbar clusters turn out to be similar to the surveyed ones. A very strange implication of this procedure is that, in effect, the IFHS ignores the all of its Baghdad and Anbar data. The governorate-level estimates are completely determined by the IBC/IFHS ratio for the six high mortality governorates and the IBC numbers for Baghdad and Anbar.

[33] Conversely, if IBC's coverage is worse in Baghdad than it is in the six high-mortality governorates than the IFHS adjustment would be too small.

a similar adjustment for the other surveys. This means that when comparing estimates across surveys one should either compare unadjusted estimates with unadjusted estimates or else compare adjusted estimates with adjusted estimates. Unfortunately, this point has often been overlooked in discussions of Iraq estimates and adjusted IFHS estimates are often compared to unadjusted estimates from other surveys.[34]

The second problem is that the adjustment itself is ill-motivated and not grounded in any statistical procedure. The sole motivation IFHS (2008a) offers for its upward adjustment is "household dissolution after the death of a household member." The idea seems to be that, for example, there is a death in a household "A" which then merges with a household "B" to form a new household which we will call "AB". A respondent representing household AB, but who was originally in household B, might then fail to report the death that occurred in household A before households A and B merged. This is, indeed, a possible scenario. However, it is also possible that after a death in household A the remaining members might split apart with some joining a household B and others joining a household C.[35] In this case, the death from household A could get reported either by merged household AB or by merged household AC. Thus, the original death from household A can easily get too much weight in the sample rather than too little weight as assumed in IHFS (2008a). Therefore, the motivation that IHFS (2008a) gives for its upward adjustment actually cuts in opposite directions and it is unclear that household dissolution really is a source of downward bias even. This, however, is only part of the problem with this adjustment. If IFHS (2008a) wished to adjust for household dissolution it should have modeled the phenomenon and applied a statistical correction procedure rooted in data. Instead, IFHS (2008a) simply scaled up its estimate by an arbitrarily chosen 54% (with arbitrary specified uncertainty around this scale-up factor added in as well). This means that roughly 50,000 out of 151,000 violent deaths in the IFHS (2008a) estimate have little basis in data or in statistical methods.

4. Conclusion

This article identifies a number of factors to consider in evaluating conflict surveys, including sampling procedures, mechanisms for ensuring the integrity of the data collection process, the appropriateness of extrapolations and the setting of baselines in the case of excess death calculations (if these are to be done). Published descriptions of methods are sometimes vague or ambiguous, but when they are relatively clear they often point towards significant weaknesses. These include sampling biases, big extrapolations from relatively small surveyed populations to much larger affected

[34] Note that, because of this confusion, in Table 1 I report both adjusted and unadjusted figures for the IFHS. It would be correct to report only unadjusted figures. However, doing this would confuse some readers who are used to seeing only adjusted figures for the IFHS.

[35] Note that the average household size in Iraq is 6.4 according to the IFHS so if one person dies there would be five survivors on average. It is hard to absorb five people into a single household so it is plausible that many households would split into separate pieces if they dissolve.

populations, unclear incentives of interviewers or interviewees, supervision methods that may be inadequate to combat these incentives and inadequate acknowledgement of the uncertainties underlying estimates that are made. Unfortunately, quality evaluations of conflict surveys are often hampered by researchers' failures to disclose their methods. Inadequate disclosure of methods, whether through negligence or stonewalling, goes against scientific principles and the public trust, and should not be tolerated by the community.

Gross overestimates of war deaths in the DRC made by the IRC, and in Iraq made by the L2 and ORB surveys, have shaken confidence in the survey approach to measuring war deaths. However, analysis of these studies shows that they all have clearly identifiable faults that point toward what went wrong. This is good news. If it were impossible to discern when an estimate is extremely far from the truth then the survey approach would become unviable.

There have been some successes in the literature. The Spiegel and Salama (2000) estimate and seems to have generated an important insight into the age distribution of war victims. The Depoortere et al. (2004) study made plain the severity of the Darfur crisis without making inappropriate extrapolations.

Far more research is needed on survey methodologies for measuring war deaths to illuminate their limits and potential. Some knowledge needs to be developed mostly from scratch, such as the performance and validation of various sampling schemes for estimating violent deaths. At the same time, some relevant knowledge, such as large literatures on survey quality (e.g. Biemer and Lyberg, 2003) and on doing surveys in developing countries (e.g., Asher, 2009), is currently sitting on the shelf largely unnoticed and could be absorbed into the conflict studies field with relative ease.

Bibliography

American Association for Public Opinion Research (AAPOR) (2006) AAPOR code of professional ethics and practice. http://www.aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/AAPORCodeofProfessionalEthicsampPractice/default.htm.

American Association for Public Opinion Research (AAPOR) (2009a) AAPOR Finds Gilbert Burnham in Violation of Ethics Code. http://www.aapor.org/AAPOR_Finds_Gilbert_Burnham_in_Violation_of_Ethics_Code/1383.htm.

American Association for Public Opinion Research (AAPOR) (2009b) AAPOR Releases Additional Detail on AAPOR Standards Violation. http://www.aapor.org/uploads/AAPOR_Press_Releases/BurhnamDetailWebsite.pdf.

American Association for Public Opinion Research (AAPOR) and American Statistical Association (ASA) (2003) Interviewer fabrication in survey research. http://www.amstat.org/sections/SRMS/fabrication.pdf.

Ali, M. and Shah, I. (2000) Sanctions and Childhood Mortality in Iraq. *Lancet* **355**(9218), 1851-93.

Ali, M., Blacker, J. and Jones, G. (2003) Annual Mortality Rates and Excess Deaths of Children under Five in Iraq 1991-98. *Population Studies* **57**(2), 217-26.

Asher, J. (2009) Developing and Using Surveys to Estimate Casualties Post-Conflict: Developments for the Developing World, paper presented at International Conference on Recording and Estimation of Casualties, Carnegie Mellon University and University of Pittsburgh.

Asher, J., Scheuren, F., and Banks, D. eds. (2008) *Statistical Methods for Human Rights*. Dordrecht, The Netherlands: Springer.

Biemer, P. and Lyberg, L. (2003) *Introduction to Survey Quality.* Hoboken New Jersey: Wiley.

Bloomberg School of Public Health (2009) Review Completed of 2006 Iraq Mortality Study.
http://www.jhsph.edu/publichealthnews/press_releases/2009/iraq_review.html.

Bredl, S., Winker, P. and Kötschau, K. (2008) A Statistical Approach to Detect Cheating Interviewers. working paper 39, Zentrum für Internationale Entwicklungs, Universität Gieβen.

Brunborg, H., Tabeau, E. and Urdal, H. eds. (2006) *The Demography of Armed Conflict*. Dordrecht, The Netherlands: Springer.

Burnham, G, Lafta, R., Doocy, S. and Roberts, L. (2006) Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey, *Lancet* **368**(9545), 1421-1428.

Coghlan, B. et al. (2006) Mortality in the Democratic Republic of Congo. *Lancet* **367**(9504), 44-51.

Coghlan, B. et al. (2008) Mortality in the Democratic Republic of Congo: An Ongoing Crisis. New York: International Rescue Committee,
http://www.theirc.org/sites/default/files/migrated/resources/2007/2006-7_congomortalitysurvey.pdf.

Coghlan et al. (2009) Update on Mortality in the Democratic Republic of Congo: Results From a Third Nationwide Survey, *Disaster Medicine and Public Health Preparedness*, **3**(2), 88-96

Daponte, B. (2007) Wartime estimates of Iraqi civilian casualties. *International Review of the Red Cross*, **89**(868), 943-957.

Degomme, O. and Guha-Sapir, D. (2010) Patterns of Mortality in the Darfur Conflict. *Lancet* **375**(9711), 294-300.

Depoortere, E. et al. (2004) Violence and mortality in west Darfur, Sudan: epidemiological evidence from four surveys. *Lancet* **364**(9442), 1315-1320.

Dyson, T. (2009) New Evidence on Child Mortality in Iraq. *Economic and Political Weekly* **44**(2), 56-59.

General Accounting Office (2006) *Darfur Crisis: Death Estimates Demonstrate Severity of Crisis, but Their Accuracy and Credibility Could be Enhanced*. Washington DC: GAO-07-24 Darfur.

Grais, R., Rose, A.M.C., and Guthmann, J.P. (2007) Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. *Emerging Themes in Epidemiology*. **4**(8).

Hagan, J. and Palloni, A., (2006) Social Science: Death in Darfur, *Science* **313**(5793), 1578-1579.

Hicks, M.J., et al. (2009) The Weapons that Kill Civilians. *New England Journal of Medicine* **369**(16), 1585-1588.

Hill, K. and Choi, Y. (2006) Neonatal Mortality in the Developing World. *Demographic Research* **14**(18), 429-452.

Iraq Family Health Survey Study Group (IFHS) (2008a) Violence-Related Mortality in Iraq from 2002 to 2006. *New England Journal of Medicine* **358**(5), 484-493.

Iraq Family Health Survey (2008b) IFHS web site. http://www.emro.who.int/iraq/surveys_ifhs.htm.

Iraq Living Conditions Survey 2004 (ILCS) (2005a) Overview. http://www.wadinet.de/news/dokus/livingconditions2004.htm.

Iraq Living Conditions Survey 2004 (ILCS) (2005b) Volume I: tabulation report. http://www.fafo.no/ais/middeast/iraq/imira/Tabulation%20reports/tabulation%20eng.pdf.

Johnson, N., Spagat, M., Gourley, S., Onnela, J. and Reinert, G. (2008) Bias in epidemiological studies of conflict mortality. *Journal of Peace Research* **45**(5), 653–664.

Kalyvas, S. (2006) *The Logic of Violence in Civil Wars*. Cambridge, UK: Cambridge University Press.

Laaksonen, S. (2008) Retrospective Two-Stage Cluster Sampling for Mortality in Iraq. *International Journal of Market Research* **50**(3) 403–417.

Macro International Inc. (2008) Democratic Republic of the Congo: Demographic and Health Survey 2007 Key Findings. http://www.measuredhs.com/pubs/pdf/SR141/SR141.pdf.

Onnela, J.P., Johnson, N.F., Gourley, S., Reinert, G. and Spagat, M. (2009) Sampling Bias due to Structural Heterogeneity and Limited Internal Diffusion. *European Physics Letters* **85**(2), 28001.

Opinion Research Business (ORB) (2008) Update on Iraqi Casualty Data. http://www.opinion.co.uk/Newsroom_details.aspx?NewsId=120.

Pedersen, J. (2009) *Health and Conflict: A review of the links*. Oslo, Norway: Fafo-report 2009:20, http://www.fafo.no/pub/rapp/20110/20110.pdf.

Roberts, L (2000) Mortality in eastern DRC: Results from Five Mortality Surveys. New York: International Rescue Committee, http://www.smallarmssurvey.org/files/portal/issueareas/victims/Victims_pdf/2000_IRC.pdf.

Roberts, L. et al. (2001) Mortality in eastern Democratic Republic of Congo: Results from Eleven Mortality Surveys. New York: International Rescue Committee, http://www.grandslacs.net/doc/3741.pdf.

Roberts. L. et al. (2003) Mortality in the Democratic Republic of Congo: Results from a Nationwide Survey. New York: International Rescue Committee, http://www.reliefweb.int/library/documents/2003/irc-drc-8apr.pdf.

Roberts, L., Lafta, R, Garfield, R., Khudhairi, J and Burnham, G. (2004) Mortality before and after the 2003 invasion of Iraq: cluster sample survey. *Lancet*, **364**(9448), 1857-1864.

Rosenblum, M. and van der Laan, M.J. (2009) Confidence Intervals for the Population Mean Tailored to Small Sample Sizes, with Applications to Survey Sampling. *International Journal of Biostatistics* 5 (1).

SMART Methodology Version 1 (2006), http://www.smartindicators.org/.

Spagat, M. (2009a) Mainstreaming an Outlier: The Quest to Corroborate the Second Lancet Survey of Mortality in Iraq. Forthcoming in *Defense and Peace Economics*, http://personal.rhul.ac.uk/uhte/014/Mainstreaming.pdf.

Spagat, M. (2009b) The Reliability of Cluster Surveys of Conflict Mortality: Violent Deaths and Non-Violent Deaths. Presentation given at International Conference on Recording and Estimation of Casualties, Carnegie Mellon University and University of Pittsburgh, http://personal.rhul.ac.uk/uhte/014/Pittsburgh%202009.pdf.

Spagat, M. (2010) Ethical and Data-Integrity Problems in the Second *Lancet* Survey of Mortality in Iraq. *Defense and Peace Economics*, **21**(1), 1-41.

Spagat, M. and Dougherty, J. (2010) Conflict Deaths in Iraq: A Methodological Critique of the ORB Survey Estimate. *European Survey Research*, 4 (1), 3-15.

Spiegel PB and Salama P. (2000) War and mortality in Kosovo, 1998–99: an. epidemiological testimony. *Lancet* **355**(9222), 2204–09.

Sullivan, J., Mashkeev, A. and Katarbayev, A. (2000) Infant and Child Mortality. in Macro International Inc. (2000) *Kazakhstan Demographic and Health Survey 1999.* Washington, DC: Macro International Inc., http://www.measuredhs.com/pubs/pdf/FR111/09Chapter9.pdf.

Uppsala Conflict Data Program (2010) Codebooks. http://www.pcr.uu.se/research/UCDP/data_and_publications/codebooks.htm.

Zaidi, S. (1997) Child Mortality in Iraq, *Lancet* **350**(8988), 1105.

Zaidi S, Smith, Fawzi, M.C. (1995) Health of Baghdad's children. *Lancet* **346**(8995), 1485.