ETHICAL AND DATA-INTEGRITY PROBLEMS IN THE SECOND *LANCET* SURVEY OF MORTALITY IN IRAQ

Running title: Second *Lancet* Survey of Iraq

September 2008

Michael Spagat
Department of Economics
Royal Holloway College
Egham, Surrey, TW20 0EX
United Kingdom
HiCN, CEPR, WDI

Abstract

This paper considers the second *Lancet* survey of mortality in Iraq published in 2006. It presents evidence suggesting ethical violations to the survey's respondents including endangerment, privacy breaches and in obtaining informed consent**.** Breaches of minimal disclosure standards examined include non-disclosure of the survey's questionnaire, data-entry form, data matching anonymized interviewer IDs with households and sample design. The paper also presents evidence relating to data fabrication and falsification which falls into nine broad categories. This evidence suggests that this survey cannot be considered a reliable or valid contribution towards knowledge about the extent of mortality in Iraq since 2003.

JEL codes: N4, I1, C8

Key words: Iraq mortality, *Lancet* survey, conflict, ethics, fabrication, falsification

1. INTRODUCTION

More than five years have elapsed since the invasion of Iraq. The human losses suffered by the Iraqi people during this period have been staggering. It is clear that there have been many tens of thousands of violent deaths in Iraq since the invasion.[1] The Iraq Body Count project (continuously updated) has documented a minimum of 86,849 violent deaths of civilians in Iraq through the end of August of 2008.[2] Total violent deaths already must be well in excess of 100,000 once combatants, non-Iraqis (including coalition soldiers) and undocumented Iraqi deaths are added in. Iraq Family Health Survey Study Group (2008a), a recent survey published in the *New England Journal of Medicine* (hereafter the "IFHS"), estimated 151,000 violent deaths of Iraqi civilians and combatants from the beginning of the invasion until the middle of 2006.

Burnham et al. (2006a) (hereafter "L2"), a widely cited household cluster survey estimated that Iraq had suffered approximately 601,000 violent deaths, i.e. four times as many as the IFHS estimate, during almost precisely the same period as covered by the IFHS study.[3] The L2 data are also discrepant from data provided by a range of other reliable sources, most of which are broadly consistent with one another.[4] Nonetheless, there remains a widespread belief in some public and professional circles that the L2 estimate may be closer to reality than the IFHS estimate.[5]

Not least because of these strong and somewhat evidence-resistant attitudinal biases, it is important that researchers develop the best possible understanding of the large human losses in Iraq, building on reliable information and discarding unreliable information. Policy should be based on evidence rather than myth or political preferences.

This paper is a contribution towards an evidence-based approach, and outlines two linked analyses. The first analysis lays out ethical concerns in relation to the conduct of L2. The second analysis points to anomalies in the data set itself, whose origin may be traced, in whole or part, to the ethical shortcomings of the study.

Analysis 1 comprises Section 2 of this paper, and examines the conformance of L2 to a number of sections of the AAPOR Code of Professional Ethics & Practices (AAPOR, 2005) published by the American Association for Public Opinion Research (AAPOR). Section 2 is structured by reproducing in italics the pertinent sections of AAPOR (2005) and then presenting relevant evidence in relation to the conformance of L2 to that code.

---

[1] There have also been large numbers of serious injuries, kidnappings, displacements and other affronts to human security.

[2] See http://www.iraqbodycount.org/, the continuously updated web site of the Iraq Body Count Project. I use hyperlinks to web-based material for the convenience of online readers. I also provide more formal references in the bibliography.

[3] For brevity I refer to this Burnham et al. (2006) article as "L2", i.e., the second *Lancet* article on mortality in Iraq. This designation distinguishes it from "L1", i.e., Roberts et al. (2004).

[4] See section 3.6 of this paper and Spagat (2008).

[5] See, for example, Steele and Goldenberg (2008) and Burkle et al. (2008) for, respectively, journalistic and academic treatments that seem to favor the L2 estimate relative to the IFHS and all the other evidence covered in section 3.6 of this paper and in Spagat (2008).

Some of the evidence in Section 2 points toward the possibility of data fabrication and falsification in L2. In Analysis 2 (Section 3) this evidence is developed further and explored. Data fabrication refers to the creation of false data by field workers. Evidence is examined in relation to the possible fabrication of violent deaths themselves, claims of death-certificate confirmations of some deaths and non-response rates. Data falsification refers to the creation of false data by one or more of the authors of a study. Falsification includes misrepresentation and suppression of other evidence relevant to the claims of that study, something I sometimes refer to as "information falsification". The evidence relating to possible fabrication and falsification in L2 is analyzed under nine broad categories

In Section 4 the findings of the paper are summarized, and the case for a formal investigation of L2 is examined.

2. AAPOR CODE OF PROFESSIONAL ETHICS & PRACTICES

This section covers sections of the AAPOR Code, (AAPOR, 2005) that may have been violated in the order that they appear in the Code. Note that the AAPOR Code is not binding on the L2 team in any legal sense. At the same time AAPOR, and anyone else, have the right to criticize survey work that does not meet these standards.

> ***II. Principles of Professional Responsibility in Our Dealings With People***
>
> *D. The Respondent:*
>
> *1. We shall avoid practices or methods that may harm, humiliate, or seriously mislead survey respondents.*
>
> *2. We shall respect respondents' concerns about their privacy.*
>
> *3. Aside from the decennial census and a few other surveys, participation in surveys is voluntary. We shall provide all persons selected for inclusion with a description of the survey sufficient to permit them to make an informed and free decision about their participation.*
>
> *4. We shall not misrepresent our research or conduct other activities (such as sales, fund raising, or political campaigning) under the guise of conducting research.* (AAPOR, 2005)

There is evidence suggesting that the L2 authors have violated all of the above four sections of the code.[6]

The following text appears in the L2 paper:

[6] See Hicks (2006) for important background on the ethics of the L2 survey.

> "By confining the survey to a cluster of houses close to one another it was felt the benign purpose of the survey would spread quickly by word of mouth among households, thus lessoning risk to interviewers." (Burnham et al., 2006a)

Note that according to the published L2 methodology in each cluster interviews were conducted at 40 contiguous households.[7] It is, therefore, likely that word about the survey would indeed have traveled from household to household, even without special encouragement by L2 field teams. In fact, the L2 field teams actively promoted word-of-mouth explanations of the purpose of the study with local neighborhood children playing central roles in these explanations. Burnham (2007), a lecture given at MIT, elaborated on the survey's reliance on local neighborhood children to explain the purpose of the survey and spread news of its benign intent:

> "They [the interviewers] went out house to house in their white coats so that they couldn't be mistaken for being somebody else. They, first off, rounded up the children to explain what this survey was about, sent out the children to the households to explain to the neighbors what was going on and so forth, to try and reduce the risks that were involved." (Burnham, 2007, around minute 23.19)

Interviewed for Munro (2008), an article in the *National Journal*, Gilbert Burnham confirmed this use of neighborhood children and that the interviewers wore white coats.[8] He further explained that interviews were conducted on the doorsteps of respondents.

Several ethical problems ensue from conducting interviews within compact neighborhoods on contiguous groups of homes, communicating the purpose of the survey through word of mouth, relying particularly on local children to spearhead these word-of-mouth dynamics, conducting interviews on doorsteps and using interviewers clad in highly visible clothing.

A. Such procedures compromise confidentiality (II.D.2). In each locality the identities of all interviewed households would be widely known. Local residents would readily observe interviewers progressing along a sequence of connected households wearing unusual white coats. Doorstep interviews would have been visible to passers by and neighbors. Parts of interviews could have been audible to third parties. Field teams specifically encouraged spreading news of the survey through word of mouth, further eroding confidentiality. Children, not naturally discrete, were actively engaged in canvassing the neighborhood to explain the survey.

It is likely that perpetrators of violence would have sometimes been aware that relatives of their victims were being interviewed for the L2 study. In many cases perpetrators would have been local criminals or militia members who might even have been acquainted with respondents. Local militias would have learned quickly that white-coated strangers had entered their neighborhoods and had "rounded up" local children. It has been acknowledged that L2 field teams did encounter militias in the field (Burnham

[7] In practice there was some variation from the intention of conducting 40 interviews in each cluster.
[8] In Burnham and Roberts (2008) Burnham and L2 co-author Les Roberts stated that both children and adults, not just children, were used to spread word of the survey.

et al., 2006b, Appendix B). L2 attributes 31% of the violent deaths in its sample to coalition forces with the remainder blamed on "other" and "unknown" agents. This implies that respondents did discuss identities of perpetrators on their doorsteps, at least in general terms

Allowing the identities of respondents to leak into the local public domain would breach confidentiality (II.D.2). Such breaches could have been life-threatening (II.D.1), even if the precise answers given by these identified respondents were not discovered by third parties. Consider, for example, what might have happened to female respondents whose husbands had been killed by local militias if these violent groups discovered that these widows had been interviewed by a violence survey.

B. The process of obtaining informed consent for the survey was compromised by the L2 field procedures (II.D.3). The L2 field teams had no means to control how the purpose of the study was explained to potential respondents. By encouraging neighbors, with a particular emphasis on neighborhood children, to explain the purpose of the study, the field teams set in motion uncontrollable dynamics that may have distorted the perceptions of L2's potential respondents. It is no longer possible to reconstruct how individual participants, many of whom would have first learned about the study from a neighbor (adult or child), understood the purpose of the study at the moment they consented to be surveyed. Initial misimpressions may have been repaired by a consent script read before field teams obtained (oral) consent for the interviews. However, at present it is unclear whether L2 had a standard oral consent script and, if so, what its content was. The L2 authors have refused to disclose any informed consent script that might have been read to potential subjects.[9] If there was no oral consent script then any false impressions spread through word of mouth would have been left unaddressed.

There is, moreover, a sense in which L2's consent procedures, whatever these might have been, were rendered irrelevant due by the confidentiality issues discussed above. Approaches to potential respondents were essentially public events at the local level and would often have been known by local militias or criminals. A person could answer the door and refuse to be interviewed but he or she might still not be able to demonstrate to intimidating observers that he or she had truly refused. Local militia members, for example, may have simply assumed that someone who had been approached by the survey had disclosed information detrimental to the interests of the militia. Such an individual might have suffered simply from answering the door, regardless of whether or not he or she had actually consented to be interviewed.

C. Respondents may have been misled (III.D.1) and/or the research misrepresented either by L2 field-team members themselves or by adult or child neighbors of respondents, whom the field teams entrusted with explaining the purpose of the study to the local population. It would be surprising if at least some neighbors, particularly

[9] Dr. Madelyn Hicks of the Institute of Psychiatry of the University of London specifically requested oral consent scripts in English and all non-English languages used but was refused by the L2 authors (personal communication).

children, did not mislead or misrepresent the survey to some respondents. The burden must be on the authors of the study to demonstrate that this did not happen.

In addition, respondents may have been misled by L2 field-team members. According to Burnham et al. (2006a):

> "Participants were assured that no unique identifiers would be gathered." (Burnham et al., 2006a)

Yet, the following data entry form was submitted to the World Health Organization (WHO) by L2 co-author Riyadh Lafta, as the data entry form used in L2 (Munro and Canon, 2008): This form requires entries of names, clearly unique identifiers, for heads of households and for all household members who have either died or were born since 2002. If this data-entry form really was implemented in the field then it appears that unique identifiers were gathered.

Governorate Cluster No. House No. Name of householder

No. of family members Males Females

No. of live births since 2002: Name sex Date of birth
1. . ……..
2. …… ……
3. ……..

No. of deaths since 2002
Name Sex Age Date of death cause (in details):
1. ……..

2. ……...

3. ……

Presence of death certificates: Yes No

Hospitalization due to violence: Age Sex Date cause

In-migration out-migration (during that period)

I am not aware of any evidence suggesting that either the IFHS or the Iraq Living Conditions Survey (ILCS, 2005a) used children or word-of-mouth to explain their purposes or that either of these surveys compromised confidentiality by conducting interviews on doorsteps or wearing conspicuous clothing. The IFHS questionnaire, posted at IFHS (2008b), provides an informed consent script right at the beginning.

The use of children, doorstep interviews and the wearing of conspicuous clothing all probably had the effect of reducing risk to interviewers. Unfortunately, some of these risks were also probably shifted onto respondents and the children who were used. In

situations where it is actually necessary to take such measures to protect interviewers it is probably better to postpone a survey until conditions are more favorable.

III. Standards for Minimal Disclosure

….. At a minimum the following items should be disclosed.

1. Who sponsored the survey, and who conducted it. (AAPOR, 2005)

Munro and Canon (2008) revealed that the Open Society Institute of George Soros was an important funder of L2, a fact that was not disclosed in the L2 paper (III1). IFHS (2008b) discloses that the IFHS "Was financially supported by WHO core budget and the United Nations Development Group Iraq Trust Fund (European Commission)." ILCS (2005b) discloses that "The United Nations Development Program (UNDP commissioned the study with a generous grant from the Kingdom of Norway."

2. The exact wording of questions asked, including the text of any preceding instruction or explanation to the interviewer or respondents that might reasonably be expected to affect the response. (AAPOR, 2005)

The L2 authors have not publicly released their questionnaire in any language: English, Arabic or Kurdish (III2). It is not clear at this stage that there was a formal questionnaire for L2 and there is no way to know how questions were worded in the field.[10] Various researchers, such as Fritz Scheuren of NORC and Madelyn Hsiao-Rei Hicks of the Institute of Psychiatry in London, have requested copies of the L2 questionnaire and have been refused by the L2 authors (personal communications). Scheuren was also told that the questionnaire exists only in English and that L2 interviewers, said to be fluent in both Arabic and English, translated the questionnaire into Arabic in the field. Several problems ensue.

A. On-the-spot translation of questions by interviewers implies that exact wordings of questions as asked in the field would have varied from interview to interview and from interviewer to interviewer.

B. There is no indication that provisions were made for conducting interviews in Kurdish or even that any of the interviewers spoke Kurdish. If so, then it seems unlikely that all heads of households or spouses selected for interviewing by L2 could have been interviewed effectively in Arabic or English. Even if possible, it would not be best practice to interview only in Arabic or English in the Kurdish zone of Iraq.

In contrast, the questionnaires for IFHS and the ILCS were both developed in English, then translated into Arabic and two versions of Kurdish, and then back-translated into English to control translation quality.

[10] Note that the document submitted by Riyadh Lafta to the World Health Organization is really a data-entry form and not a questionnaire. It does not give any wordings of questions, exact or otherwise.

Iraq Mortality Survey Template

(After reading the consent statement, you should ask permission and record if the household provides consent.)

1) Who lives in this household? (Resident means spent most of the past 3 months sleeping in this household.) (only record M/F and the age, if less than 4 years, record age in months)

2) Have your family lived in this household since Jan. 1, 2002? (If no, obtain details. Only record deaths from elsewhere if majority of old family members are here now.)

3) Has any member of the household been born since Jan. 1, 2002? (record date)

4) Has any member of the household died since Jan. 1, 2002? (If yes, record Age, Gender, Date of death, Cause of death)

5) Did anyone else live here for part of this time or was one of these individuals away for more than 3 months during this period ?

(Thank them for their cooperation.)

Mortality Survey data form

Cluster #_____ Date_____ Interviewer_____

M	F	Births / deaths / missing / visitors

For decedents:

Age/gender	Date of death	Cause of death
_____	_____	_____
_____	_____	_____
_____	_____	_____

M	F	Births / deaths / missing / visitors

Munro and Canon (2008) obtained the above English-language list of questions and a data entry form from a third party who had apparently obtained it from an L2 author. However, Gilbert Burnham, Les Roberts and officials from the Bloomberg School of Public Health have declined to either confirm or deny that either of these forms was actually used in L2 or to provide the actual forms (Munro, 2008).

The "Mortality Survey data form" does not match the data entry form submitted by Riyadh Lafta to the WHO. Lafta did not submit a questionnaire to the WHO so the "Iraq Mortality Survey Template" could potentially fill this void. However, this questionnaire does not fit well with either the Lafta data-entry or the "Mortality Survey data form". For example, the "Iraq Mortality Survey Template" does not instruct interviewers to ask for death certificates when households report deaths but the Lafta data-entry form has a tick box for death certificates. The "Iraq Mortality Survey Template" instructs interviewers to record the ages of all household members yet neither of the two circulating data-entry forms contains space to record such an answer and it has been confirmed that the L2 survey did not record ages or genders of living household members.[11] Burnham et al. (2006a) states that "Deaths were recorded only if the decedent had lived in the household continuously for 3 months before the event" but the "Iraq Mortality Survey Template" requires that residents need only sleep within a household for "*most* of the past 3 months" [emphasis added]. Note also that this questionnaire mixes the terms "family" and "household" which, if done in the field, might encourage some respondents to report deaths of extended family members.

Summary: The "exact wordings of questions asked" for L2 are still unknown and may be unknowable (III.2). We cannot rule out the following possibilities:

1. There is no questionnaire in English, Arabic, or Kurdish. If there is a questionnaire then it is a puzzle why the L2 authors do not simply release it into the public domain.

2. There is a questionnaire in English but it has not been translated into Arabic or Kurdish. In this case, exact wordings of questions would have been improvised by a variety of different interviewers and would have varied from household to household. It would be impossible to reconstruct exact wordings of questions at this point in time.

It is also unclear what data entry-form was used since there are presently two competing ones in circulation.

The IFHS and ILCS questionnaires are both available in English and in Arabic at IFHS (2008b) and Fafo (undated) respectively.

Note that the ILCS and IFHS questionnaires show clearly that these surveys, in contrast to L2, both recorded household rosters, including lists of all the members of each household in their samples with gender and age information for each individual. L2's failure to record household rosters is a shortcoming according to two recent attempts to

[11] See the section labeled "corrections" of Deltoidblog (2006 and 2008).

codify and raise standards in conflict mortality surveys. The SMART Methodology states:

> "Sometimes the respondent is simply asked to state how many people are in the household. Although this is quicker, it is much less accurate than asking the respondent to list all household members. We recommend that the household members be enumerated." (SMART, 2006, p. 75)

London School of Hygiene and Tropical Medicine (undated) advises:

> "Do not just ask the respondent how many people live in the household and how many have died. You may get inaccurate or intentionally distorted responses." (LSHTM, undated, p. 109)

> *III.3. A definition of the population under study, and a description of the sampling frame used to identify this population.*
>
> *III.4. A description of the sample design, giving a clear indication of the method by which the respondents were selected by the researcher, or whether the respondents were entirely self-selected.* (AAPOR, 2005)

The authors of L2 have still not fully disclosed their sample design (Bohannon, 2008, Spagat, 2007). Gilbert Burnham and Les Roberts have stated frequently that the L2 field teams did not follow the sampling methodology that was published in the *Lancet* but they have not supplied a viable alternative. Burnham and Roberts have also issued a series of contradictory statements about their sampling procedures and have either destroyed or not collected evidence necessary to evaluate these procedures.

Johnson et al. (2008) suggests that sampling procedures described in L2 might have caused substantial upward bias in L2's estimate of the number of violent deaths. This idea is based on L2's published description of the final stages of its sampling methodology:

> "The third stage consisted of random selection of a main street within the administrative unit from a list of all main streets. A residential street was then randomly selected from a list of residential streets crossing the main street." (Burnham et al., 2006a)

The published description goes on to explain that the field teams would then select a household on this residential cross street to a main street and then conduct interviews at forty contiguous households.

Johnson et al. (2008) argues that residential cross streets to main streets would suffer from higher-than-average violence within the context of the Iraq war because:

a. Crowded markets, cafes restaurants and other attractions will be on such streets.

b. Military patrols focus on such streets. In fact, many military vehicles can only go down the larger streets.

c. Abductions and mass shootings also tend to be on such streets. For example, Sunnis would not travel deep into Shiite territory, abduct some people and make a long drive to reach safe territory. Rather, they would make a quick foray in and out of enemy territory, perhaps just crossing over a main street that divides the two areas, and continuing only until they were just inside of a residential area.

It is, at least, plausible that such a bias could exist and that it could be substantial. In the present article I do not focus directly on the potential size of this possible bias. Rather, I consider the responses of the L2 authors to the suggestion of possible sampling bias in L2.

This picture below illustrates the types of areas that will be missed by a methodology of conducting interviews at forty contiguous households beginning at a household on a residential cross street to a main street. Scope is limited for reaching areas not actually on residential cross streets to main streets.

Quoting again from L2:

> "The third stage consisted of random selection of a main street within the administrative unit from a list of all main streets." (Burnham et al., 2006a, emphasis added).

These lists of main streets are at the core of the claimed sampling methodology. Yet, the L2 authors have refused to provide these lists or even clarify where they came from.[12] Without this information we cannot assess the sampling frame for the study (III.3) and we cannot know the sample design fully (III.4).

Gilbert Burnham did make aspects of the sampling methodology fairly concrete in Biever (2007), an interview with the *New Scientist*.

> "The interviewers wrote the principal streets in a cluster on pieces of paper and randomly selected one. They walked down that street, wrote down the surrounding residential streets and randomly picked one. Finally, they walked down the selected street, numbered the houses and used a random number table to pick one. That was our starting house, and the interviewers knocked on doors until they'd surveyed 40 households…. The team took care to destroy the pieces of paper which could have identified households if interviewers were searched at checkpoints." (Biever, 2007, emphasis added.)

Whatever its strengths or weaknesses, this does seem to be a procedure that can be followed in the field. The L2 authors may no longer be able to specify their sample design since these pieces of paper have been destroyed. But they should be able to supply lists of principal streets or at least specify how many such streets there were per governorate.

Burnham explains that the sampling information was destroyed to protect the identities of respondents, but this explanation is inadequate. Pieces of paper with lists of principal streets and surrounding streets would be of no use for identifying households included in the survey. Even lists of all of the households on a street that was actually sampled would not be usable for identifying particular L2 respondents. On the other hand, the L2 data-entry form that Riyadh Lafta submitted to the WHO contains spaces for listing the name of each head of household in addition to names of people who died or were born during the L2 sampling period. If the field teams could travel around with pieces of paper containing the names of their respondents plus many of their family members then they did not have to destroy lists of streets. Finally, as noted above in section 2, the lists of L2's respondents would have been widely known at the local level in any case.

The L2 authors have often dismissed the possibility of sampling bias by stating that they did not actually follow the sampling procedures that they claimed to have followed in their *Lancet* publication. For example, Burnham and Roberts (2006a) write that they had removed the following sentence from their description of their sampling methodology at the suggestion of peer reviewers and the editorial staff at the *Lancet*:

[12] For example, Seppo Laaksonen, a professor of survey methodology in Helsinki, requested and was denied any information on main streets, even the average number of main streets per cluster (Laaksonen, 2008).

> "As far as selection of the start houses, in areas where there were residential streets that did not cross the main avenues in the area selected, these were included in the random street selection process, in an effort to reduce the selection bias that more busy streets would have." (Burnham and Roberts, 2006a)

Thus, this part of the description of sampling methodology should have read:

> "The third stage consisted of random selection of a main street within the administrative unit from a list of all main streets. A residential street was then randomly selected from a list of residential streets crossing the main street. *As far as selection of the start houses, in areas where there were residential streets that did not cross the main avenues in the area selected, these were included in the random street selection process, in an effort to reduce the selection bias that more busy streets would have.*" (Original text from Burnham et al. (2008) with new text italicized)

Combining this with Gilbert Burnham's *New Scientist* interview already quoted (Biever, 2007) would imply that at each location:

A. Field teams wrote names of main streets on pieces of paper and selected one street at random.

B. The field teams then walked down this street writing down names of cross streets on pieces of paper and selected one of these at random.

C. The field teams then became aware of all other streets in the area that did not cross the main avenues and may have selected one of these instead of one of the cross streets written on pieces of paper. This wide selection was done according to an undisclosed procedure.

The Biever (2007) description of Burnham does outline a sampling procedure that could have been followed and is broadly consistent with the published methodology. If other types of streets, beyond those that would be covered by the published methodology, were included in the sampling procedures then the authors need to specify how these streets were included. More fundamentally, how did the field teams discover the existence of such streets that could not be seen by walking down principal streets as described by Burnham in Biever (2007)? The L2 field teams would not have brought detailed street maps with them into each selected area or else it would not have been necessary to walk down selected principal streets writing down names of surrounding streets on pieces of paper. We can also rule out the possibility that the teams completely canvassed entire neighborhoods and built up detailed street maps from scratch in each location. Developing such detailed street maps would have been very time consuming and the L2 field teams had to follow an extremely compressed schedule that required them to perform forty interviews in a day (Hicks, 2006).

In Giles (2007), an article in *Nature*, Burnham and Roberts suggested one possible explanation on how the field teams had managed to augment their street lists beyond streets that could be seen by walking down a main street but this suggestion was rejected by an L2 field-team member interviewed by *Nature*:

> "But again, details are unclear. Roberts and Gilbert Burnham, also at Johns Hopkins, say local people were asked to identify pockets of homes away from the centre; the Iraqi interviewer says the team never worked with locals on this issue." (Giles, 2007)

Even if locals had identified such "pockets of homes away from the centre" the authors still would have to specify how these were included in the randomization procedures. Indeed, involving local residents in selecting the streets to be sampled would seem to be at odds with random selection of households. Locals could, for example, lead the survey teams to particularly violent areas.

Burnham and Roberts have induced further confusion about their sample design by issuing a series of contradictory statements.

> "The sites were selected entirely at random, so all households had an equal chance of being included." (Burnham et al, 2006b, emphasis added)

> "Our study team worked very hard to ensure that our sample households were selected at random. We set up rigorous guidelines and methods so that any street block within our chosen village had an equal chance of being selected." (Burnham and Roberts, 2006b, emphasis added)

> "… we had an equal chance of picking a main street as a back street." (The National Interest, 2006).

These statements contradict each other and the methodology published in the *Lancet*. Some streets are much longer than others. Some streets are much more densely populated than others. Such varied units cannot all have equal probability of selection. If, for example, every street block had an equal chance of selection then households on densely populated street blocks would have lower selection probabilities than households on sparsely populated street block. If main streets are more densely populated on average than back streets are and main streets and back streets have equal selection probabilities then households on main streets would have lower selection probabilities than households on back streets.

Thus, the L2 survey appears to violate standards III.3 and III.4 of the AAPOR Code of Professional Ethics and Practices.

The sampling methods for the ILCS are explained briefly in ILCS (2005a) and in great detail in ILCS (2005b, Appendix 2). The IFHS sampling methods are explained in IFHS (2008a), including in the supplementary appendix. The sampling methods have been well disclosed for these surveys.

> *III.5. Sample sizes and, where appropriate, eligibility criteria, screening procedures, and response rates computed according to AAPOR Standard Definitions. At a minimum, a summary or disposition of sample cases should be provided so that response rates could be computed.* (AAPOR, 2005)

L2 does give information on response rates but this information is unlikely to be correct. L2 reports nobody home in 16 households out of 1849 (0.9%) and refusals to participate

from 15 households (0.8%) This degree of success seems especially unlikely given the rushed conditions under which the survey was conducted with field teams regularly conducting 40 interviews in a single day.[13] L2 methodology did not follow a common practice, employed in several recent surveys in Iraq including the IFHS and the ILCS, of making three visits to a selected household before accepting failure to make contact. For L2, a head of household or spouse had to be present and agreeable for an interview within a single time window of perhaps 20-30 minutes almost without fail with no opportunity for repeat visits. The L2 paper plus a further clarification by Gilbert Burnham also reports that its field teams conducted interviews in 52 clusters and that there was only one security-related failure to reach a selected cluster, which was in the governorate of Wasit.[14]

The IFHS gives a rather direct comparison with L2 since the IFHS field work was conducted only a few months after the L2 field work. The IFHS failed to visit 115 out of its 1,086 clusters (10.6%) due to security reasons. These problems encountered by IFHS field workers cast doubt on the L2 report of only one failed cluster visit in 52 attempts (1.9%) due to security reasons. Assume that the IFHS success rate in cluster visits (89.4%) is the true rate for L2 and that the results of attempted visits (success or failure) are statistically independent across these attempts. Then the odds against 0 or 1 failed visits out of 52 attempts would be 47 to 1.

The IFHS disaggregates its success rates in visiting clusters by governorate: 34.2% (37/108) for Al-Anbar, 67.7% (65/96) for Baghdad, 83.3% (60/72) for Nineveh and 98.1% (53/54) for Wasit. If we take these percentages as the true ones for L2 and again assume independence across visits then the odds against the record of L2 in Baghdad, 12 successes in 12 attempts, are 108 to 1 against. The odds against L2's 5 successes in five attempts in both Al-Anbar and Nineveh are, respectively, 214 to 1 and 2.5 to 1 against.[15] The compound odds against 22 successful cluster visits in 22 attempts in these three insecure governorates are 57,780 to 1 against. Somewhat strangely, Wasit was the only governorate for which L2 reported a security-related failed cluster visit although the IFHS experience of 53 successes in 54 attempts suggests that such a failure would be improbable.

For clusters actually visited the IFHS failed to make contact 3.4% of the time compared to L2's rate of 0.9%. Assuming independence across visits and a success probability of

[13] Again, see Hicks (2006). It is claimed that one field team of four would divide into two sub-teams of two, each conducting approximately 20 interviews in a day.

[14] Burnham et al. (2006a) reports conducting interviews at 50 clusters although results from three of the 50 were discarded for various reasons. In addition, Burnham (2007, minute 20) reports that interviews were conducted at 5 clusters in Anbar governorate, 3 of which were in Falluja, but two of these Fallujah clusters were discarded. There were, therefore, 52 clusters finished although the results in the paper are based on 47 of these clusters. At hour 1, minute 8 and 40 seconds Burnham (2007) clarifies that the only security-related failure to visit a selected cluster in L2 was in the governorate of Wasit.

[15] If we ignore Gilbert Burnham's clarification that L2 did 5 clusters in Anbar and just consider the 3 clusters that were reported in the paper then the odd against L2's success rate in Anbar would become 12 to 1 against.

96.6% for each visit, as suggested by the IFHS record, the odds against the L2 report of only 16 failed contact attempts would be more than 500,000 to 1 against.

Note that the IFHS did not give up on making contact before making three contact attempts. L2, on the other hand, had a compressed work schedule and could not have tried as hard as the IFHS did to make contact. Thus, the IFHS would have been expected to have a substantially lower no-contact rate than L2's – just the opposite of what was reported by the two surveys.

L1 (Roberts et al., 2004) was conducted by many of the same people who did L2 and the two studies shared many methodological commonalities, including strong time pressure on the field teams. L1 is, therefore, a good survey to compare with L2. On the other hand, L1 was conducted nearly two years before L2 was done. During the period in between the two surveys a large number of Iraqis were displaced with at least several hundred thousand fleeing abroad. One would expect the not-at-home rate to be higher in 2006 than it was in 2004. Yet L1 reported 64 out of 988 households visited were empty (6.5%).[16] Thus, the no-contact rate for L2 was lower by more than a factor of 7 compared to L1's. If, again, we assume statistical independence across contact attempts and that the L1 no-contact rate of 6.5% applied during the L2 period then the odds against the L2 contact record would be about $7x10^{14}$. In fact, we would have to lower the true L1 no-contact rate from 6.5% to about 1.5%, to even reduce the odds against the reported L2 rate to about 90 to 1.

The ILCS, done in 2004 like L1, reports an overall failure-to-interview-rate, mixing no-contact with refusals, of 1.6%, which is slightly lower than L2's 1.7%. There are, however, two reasons why we must adjust the ILCS rate upward in order to make an appropriate comparison with L2. First, the ILCS made three contact attempts and failed to complete interviews 2.6% of the time on its first attempts. Second, the ILCS expended considerable effort preparing the ground before selecting and contacting households. Specifically, the ILCS teams completely enumerated all the households in each cluster before selecting the particular households to be interviewed. During these enumerations field teams eliminated all housing units the teams determined to be empty.[17] Thus, L2's 1.7% failure-to-interview rate should be compared to the ILCS's 2.6% plus some upward adjustment for the percent of unoccupied housing in 2006. The field work for the IFHS was conducted only a few months after L2's field work and reported that for 0.8% of its selected households the "entire household was absent for [an] extended period" and 1.3% of the time the "dwelling [was] vacant or address not a dwelling." With an empty-housing adjustment of 2% for the ILCS, an appropriate failure-to-interview rate would be 4.6% for the ILCS compared to 1.7% for L2. Even without this adjustment the odds against the reported L2 experience, using the same methods as before, are 190 to 1. If we add in the adjustment then the odds against the L2 claim rise to nearly 100,000 to 1.

[16] L1 also reported that 5 people refused interviews (0.5%). Very low refusal rates do seem to be common features of surveys in Iraq.

[17] Personal communication with Kristen Dallen of Fafo in Norway who was closely involved in the ILCS field work.

A recent poll by ABC and other news organizations, ABC (2007a), experienced a no-contact rate of 7% and a refusal rate 35% (ABC, 2007b). It appears that the refusal rate is not strictly comparable to L2's because use of the "next-birthday" method by the ABC poll probably made it harder to progress to a successful interview for this poll than it was for L2.[18] On the other hand, the L2 methodology only allows interviews with heads of households or their spouses so some adults who might have been at home when L2 interviewers visited would have been ineligible to respond to the survey. Even if we reduce the 7% rate reported by ABC by a factor of 4 the odds against the L2 record would still remain at 934 to 1.

A recent poll (ORB, 2008) failed to interview (at least on their mortality question) 251 out of 2,414 individuals contacted (10.4%), again suggesting that the claimed L2 success rate is implausible.

To summarize, these comparisons provide evidence of fabrication and falsification both in L2's reported success rates in visiting selected clusters and in L2's reported contact rates with selected households.

As a final point on disclosure I note that an incomplete L2 dataset has been released but only selectively to certain researchers (Kaiser, 2007). Below is the key part of the data disclosure policy of the L2 researchers (Bloomberg School of Public Health, 2007).

"Conditions for the Release of Data from the 2006 Iraq Mortality Study

These data will be released on request to recognized academic institutions or scientific groups with biostatistical and epidemiological analytic capacity.

1. The data will be provided to organizations or groups without publicly stated views that would cause doubt about their objectivity in analyzing the data.

2. The data will remain the property of Johns Hopkins Bloomberg School of Public Health, and will be provided only on condition that the datasets are not shared with others.

3. Results from reanalysis of the data can be freely published in the scientific and lay press. The Johns Hopkins authors request a copy of any papers accepted for publication, for information purposes only."

(Bloomberg School of Public Health, 2007)

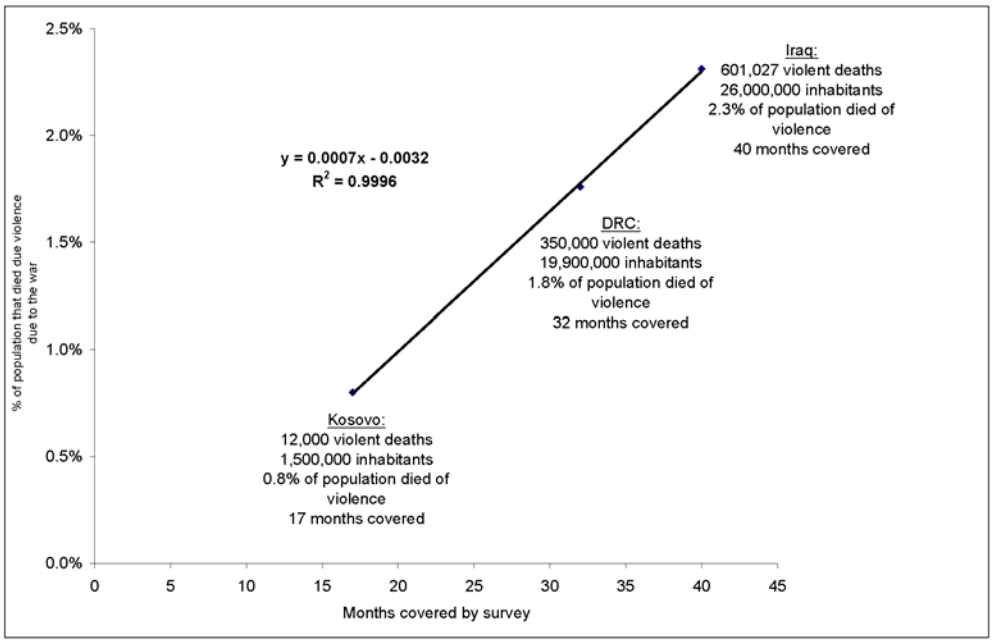
The IFHS dataset has not yet been released. The ILCS dataset is obtainable by approaching COSIT, the Iraqi statistical office, although it is not easy to obtain.

[18] For the ABC poll it appears that if the household member who will be the first to have a birthday after the date of the poll's visit could not be found or did not consent to be interviewed then the poll could not substitute another household member for the original one.

3. THE POSSIBILITIES OF FABRICATION AND FALSIFICATION IN L2

In this section I discuss a varied body of evidence for fabrication and falsification in the L2 data and paper and reports of L2 results. I have already presented some of this evidence in Section 2. I stress the evidence of fabrication/falsification in response rates and in success rates in visiting selected clusters and failure to properly disclose many aspects of the study including wordings of questions, the data-entry form, the sample design and data that matches anonymized interviewer IDs with particular interviews. In the next sub-section I take a different tack, looking at evidence for falsification by the extrapolation of L2's results from two previous studies. The main exhibit for this conclusion is the following graphic.

3.1. Evidence of extrapolation of the L2 results from previous studies



The above graphic shows results from three mortality surveys.[19] The first is the Kosovo study of Spiegel and Salama (2000). This paper is cited in Roberts et al. (2004),

[19] This graphic was passed to me by researchers who asked to remain anonymous. I have verified that the results are true and it is easy for anyone to verify the same thing.

Burnham et al. (2006a) and Burnham et al. (2006b). This is, thus, a paper that the L2 authors know well.

There was an exchange of letters in the *Lancet* of January 13, 2007. Guha-Sapir, Degomme and Pedersen (2007) questioned the L2 finding that roughly 90% of all excess deaths in Iraq were violent, contrary to findings in other war studies such as those done on the Democratic Republic of Congo (DRC). The L2 authors responded:

> "We feel a better comparison would be to the data collected during that war which showed that 1.8% of the 19.9 million people in the eastern part of the country died of violence in the first 33 months of the conflict, a proportion similar to that measured in Iraq." Burnham et al. (2007)[20]

To back up this claim they cite Roberts et al. (2001), a study of the DRC. This is the second point in the above graphic.

The third and final data point is L2 itself.

What is highly suspicious is that the three studies are in near-perfect alignment. A regression line drawn through them has an Rsquared of 0.9996. One could make slightly different assumption and feed in slightly different numbers but under any plausible scenario the fit is nearly perfect with an Rsquared of at least 0.99. All of these studies have quite large confidence intervals so the chances of their central estimates lining up so well would appear to be very small.

The Kosovo and DRC studies were in the literature for a number of years before L2 was done. Draw a line between these first two central estimates and the slope suggest that an additional 15 months of conflict will result in deaths of an additional 1% of the population. Extending the line, the 8 months by which the L2 period exceeds the DRC period would bring the total percent killed during the L2 period to just over 2.3. The fact that the L2 authors cite the DRC study as being similar to L2 in terms of the number of months and percent of population killed and the fact that the L2 authors are well aware of the Kosovo study reinforces the relevance of the graph.[21]

Professor Mark van der Laan of the University of California Berkeley quantified the probability of the three points lining up the way they do due to pure chance as 0.036.

---

[20] Note that the letter to the *Lancet* states that the DRC study covered a 33-month period. Yet the introduction to the paper the letter refers to states, correctly, that the coverage period is 32-months. Later the same paper lapses into referring to a 33-month period. The graphic in this section uses 32 months since this is the correct figure. However, the graphic barely changes if we switch to 33 months. For example, the R Squared decreases only from 0.9996 to 0.9992.

[21] Using a population estimate of 27 million rather than 26 million slightly reduces the percent of population violently killed to 2.2% and also slightly reduces the Rsquared for the regression to 0.9906. L2 used an estimate of about 27 million for the total population of Iraq and 26 million for the population actually covered by the survey, the difference being due to accidental non-coverage of Wasit governorate. The summary to L2 reports that excess deaths, violent plus non-violent, were estimated to be just over 650,000, a number which is also presented to be 2.5% of the population. 650,000 is precisely 2.5% of 26 million but is only 2.4% of 27 million. So it is clear that the L2 authors were thinking in terms of a population of 26 million.

This is based on a simulation taking 100,000 draws of three points with normal distributions and respective means and standard errors of (0.8, 0.21), (1.8, 0.4) and (2.3, 0.4) where the standard errors are suggested by the published studies (R code available upon request). Thus, this three-point diagram provides statistical evidence of data falsification although it is not definitive; we reject the hypothesis that the alignment arose by chance at the 5% level but not at the 1% level.

3.2. Risk factors for interviewer fabrication

AAPOR and ASA (2003), a joint document of AAPOR and the American Statistical Association (ASA), lists a number of risk factors for data fabrication by interviewers. Most of them are present in L2. Here is the list of risk factors with commentary on their relationship to L2.

a. hiring and training practices that ignore fabrication threats,

I am not aware of any information concerning hiring practices for L2. L2 states that the interviewers were all medical doctors with "previous survey experience and community medicine experience and were fluent in English and Arabic" but does not explain how they were hired. L2 further states that there was a 2-day training session for the field workers but the L2 researchers have refused to disclose any information on the content of these sessions other than that interviewers were "trained in the use of the questionnaire" (Burnham et al., 2006b). There is no evidence of any attention to fabrication threats in any training or hiring practices.

I have no information on hiring practices for the ILCS or the IFHS. ILCS (2005b) and IFHS (2008a, supplementary appendix) are clear that training and field testing for both surveys were extensive although they contain no information on the content of the training.

b. inadequate supervision,

None of the US-based authors were in Iraq when the field work was conducted so none of them could have provided meaningful supervision. Burnham et al. (2006a) does not claim that the US-based authors did supply any field supervision. The paper simply states that Riyadh Lafta was the field manager and supervisor. There is no information on how Lafta discharged these duties. Moreover, Lafta is not available to answer questions about how he supervised the L2 field work. He has a policy of not responding to any questions from journalists and his only interactions with researchers on this subject of which I am aware were an off-the-record meeting at the WHO at which he submitted his data-entry form. The US-based L2 researchers do not facilitate contacts with Riyadh Lafta (Munro and Canon, 2008).

The IFHS employed 112 2-person (male-female) interview teams and 100 supervisors: 21 central, 20 local and 59 in the field (IFHS, 2008b). The ILCS had five-person interview teams, each with its own supervisor (ILCS, 2005a) with additional supervision and visits

from COSIT, the Iraqi statistical department, and Fafo, the Norwegian institute that was in charge of the study.

The AAPOR/ASA document discusses a number of supervisory methods that can be employed to prevent fabrication but there is no evidence that Riyadh Lafta employed any of these methods. These include:

i. Observational Methods.

This means monitoring interviews. L2 had two field teams consisting of four interviewers who are said to have divided into sub-teams of two for actual interviewing. Thus, it was possible for Riyadh Lafta to monitor up to about 25% of all the interviews. There have, however, been no indications that Lafta actually did any such monitoring.

ii. *Recontact Methods*

These methods can involve physically revisiting households that were supposed to have been interviewed or simply calling them on the telephone or writing to them through the mail. These recontacts can be used to check data that have been collected or simply to check that interviews were actually conducted. L2 did not use any recontact methods. Furthermore, the apparent destruction of records on where interviews were conducted means that recontact of households that were interviewed for L2 was never and will never be possible.

iii. Data Analysis Methods

These methods can involve the identification of suspicious patterns of particular interviewers. The L2 authors have offered no evidence that they used such methods and have refused to cooperate with other people, such as Fritz Scheuren of NORC, who have wanted to apply them. As noted above, the L2 authors refuse to release data with anonymized interviewer IDs matched to the results of interviews.

Collection and analysis of demographic information on respondents and their families is another important, and commonly used, check against fabrication. But the L2 study did not collect demographic information on households other than the number of males and the number of females contained in each one (with some omissions).

iv. Selection Procedures

The document states that “typically 5-15% of the interviews are monitored and/or recontacted.” But L2 apparently did not have any monitoring and had no recontact. Of course, field teams would have been well aware of the lack of supervision in the study and might have acted accordingly.

All of the above supervisory methods were employed by the ILCS and the IFHS. Note, in particular, that both surveys collected data matching interviews with anonymized interviewer IDs and this information is present in the ILCS dataset that has been released.

c. lack of concern about interviewer motivation

I found no evidence of concern about interviewer motivation in the L2 study. The L2 authors have not disclosed any information about their interviewers, other than the phrase quoted above under "point a". On the other hand, I also did not find evidence of concern about interviewer motivation in materials released by the ILCS or the IFHS.

d. poor quality control

I have already discussed the lack of quality control in the collection of the data. An indicator of this lack of quality control is the L2 dataset itself which has been well-documented to contain numerous errors, omissions and inconsistencies.[22] Data that are sometimes missing include household sizes (13 times), months in which deaths occurred (57 times), and the number of males and females in each household (55 times).[23] The dataset usually gives household sizes in 2002 and 2006 plus births, deaths, immigration to and emigration from the households but for 14% of all households the identity,

Household size 2006 = Household size 2002 + births – deaths + in migration – out migration

does not hold. Occasionally the identity fails by a wide margin. The L2 paper states:

> "The interviewers then asked about births, deaths, and in-migration and out-migration, and confirmed that the reported inflow and exit of residents explained the differences in composition between the start and end of the recall period."

Thus, these inconsistencies should have been filtered out in the field but often were not. Excessive workload (the next item) may be one of the reasons why these consistency checks were not routinely applied.

In L2's single cluster that was done in the governorate of Al-Tameem data are missing on the number of males and the number of females for all 40 households. This can be viewed as another quality control issue; someone should have spotted this deficiency and sent field workers back to this cluster to gather the missing data. Note, however, that field teams consisting of four people are said to have worked in groups of two. This means that one pair should have done approximately 20 of the households in the cluster with the other pair doing the other 20 households. It is a bit implausible that both teams would have separately forgotten to record the number of males and females for their entire half of the cluster. Moreover, if these pairs were actually using the data-entry forms that Riyadh Lafta submitted to the WHO it seems unlikely that they could have gone through 20 interviews without realizing that they were not filling in the box for

[22] Kane (2007) and Laaksonen (2008) both discuss the quality of the L2 dataset.
[23] The dataset gives the year in which each death occurred, never gives exact dates of deaths and usually, but not always, gives a month of death.

gender information. Thus, perhaps interviews were not really conducted as described in the Al-Tameem cluster.

I am not aware of any similar indicators of poor quality in the ILCS or IFHS mortality data.

d. excessive workload

L2 imposed an extraordinary workload on its field workers (Hicks, 2006). Field teams were routinely expected to conduct 40 interviews in a single day. Moreover, it is claimed that the two field teams completed 52 clusters (40 interviews per cluster) in just 52 days of field work. To accomplish this task the teams had to travel all over Iraq during one of the most violent periods of the conflict, encumbered by checkpoints and poor transportation infrastructure in a country that had experienced, over the last three decades, three wars and strict economic sanctions.

The IFHS had 112 interview teams conduct 9,345 interviews in 971 clusters spread over 4 months. This works out to about 2 interviews every 3 days per team on average with a team completing a cluster of 10 households roughly every two weeks on average. These teams were supported by 100 supervisors and 55 data-entry people as well. The ILCS had 500 workers but does not give a breakdown. Since the ILCS sample size was more than twice the that of the IFHS and the IFHS was largely conducted within two months it would appear that ILCS interviewers would have experienced more time pressure than IFHS interviewers. However, time pressure on L2 interviewers would have been much greater than in either the IFHS or the ILCS.

e. inadequate compensation

f. piece-rate compensation as the primary pay structures

To my knowledge there is no information available on how the field teams were compensated for L2, the ILCS or the IFHS.

g. off-site isolation of interviewers from the parent organization

The parent organization for L2 is Johns Hopkins University so there was indeed off-site isolation of interviewers from the parent organization. No one from the parent organization was present in Iraq during the L2 field work. The IFHS and ILCS did not suffer from such off-site isolation.

It appears that the group that wrote AAPOR/ASA (2003) probably did not envision interviewers working under dangerous conditions. It is clear, however, that interviewers who must risk their lives to be out in the field will be tempted to avoid these risks by fabricating data.

To summarize, most of the risk factors for fabrication identified in the AAPOR/ASA document were present in the L2 study. Some, such as excessive workload, were present, arguably, to an extreme degree. Other factors may not have been present but cannot be ruled out based on the information that is currently available. Of course, the presence of so many risk factors for fabrication does not prove that fabrication actually occurred. Nevertheless, the above discussion demonstrates that the L2 project operated virtually without defenses against fabrication. As Fritz Scheuren of NORC pointed out:

> "They failed to do any of the [routine] things to prevent fabrication." (Munro and Canon, 2008)

3.3. A claimed work schedule that seems to be impossible without ethical transgressions

The key reference on this is Hicks (2006), developing ideas that were first expressed Bohannon (2006). This paper makes concrete the many things that L2 field teams needed to accomplish at each household and argues that it is implausible that the teams could have worked on such a punishing schedule while maintaining acceptable ethical standards.

Additional factors to those covered in the Hicks paper add further grounds for skepticism that the L2 study could have been performed as claimed. The sampling routines described above would have been time consuming. At each cluster a field team needed to walk down a main street writing down names of cross streets and then select one at random. The teams would then have to have walked the length of the selected cross street enumerating all the houses on that street so that one of these could be chosen at random as the starting point. If we accept that field teams somehow included streets that were not cross streets to main streets then even more time would have to have been spent locating these other streets. In addition, traveling from cluster to cluster while navigating checkpoints along a bad system of roads, degraded by years of conflict and sanctions would also have been very time consuming as the two field teams attempted to move from cluster to cluster.

3.4. L2 estimates compared to those of other surveys[24]

In this section I compare the distribution of violent deaths nationally and by governorate in L2 with the distribution of "war-related deaths" in the ILCS (ILCS, 2005a)) and with violent deaths in the IFHS (IFHS, 2008a). I also make some use of the database of the Iraq Body Count (IBC) project.[25]

[24] For this section I have benefited enormously from information supplied to me by Gabriel Guerrero-Serdan on the Iraq Living Conditions Survey (ILCS). Also, the L2 authors refused to give the L2 data to a number of researchers including me. Thus, I had to rely on the kind cooperation of David Kane for the figures from the L2 data appearing in Section's 3.4, 3.6 and 3.7. Although he was unable to share the actual dataset with me, he did provide answers to many specific questions that I put to him about the data.

[25] Spagat (2008) makes similar comparisons, offering a somewhat different treatment.

The ILCS, supported by the United Nations Development Program in Iraq, estimated 24,000 "war-related deaths" with a 95% Confidence Interval (CI) of 18,000 to 29,000 based on field work conducted mainly between March 22, 2004 and May 25, 2004. The ILCS had a recall period of two years so it covered slightly more than a year after the invasion of Iraq and slightly less than a year before the invasion.

First, note that non-violent death rates for L2 and the ILCS are quite similar: 4.5 and 4.8 per 1,000 per year for the ILCS period respectively. L1's non-violent death rate of 5.3 per 1,000 per year is also close to the non-violent death rates for L2 and the ILCS.

But violent-death estimates diverge dramatically, L2 versus ILCS. Even taking L2 only through March 31, 2004, eight weeks before the ILCS field work was completed, the L2 central estimate exceeds the ILCS one by nearly a factor of 3 (Table 1). This becomes almost a factor of 4 if we include April and May for L2 (Table 2).

The IFHS is suitable for comparing with L2 because it includes almost exactly the same coverage period.[26] The IFHS gives a central estimate of 151,000 violent deaths with a 95% CI of 104,000 to 223,000. The central estimate of L2 for violent deaths exceeds that of the IFHS by a factor of 4 and even the bottom of the L2 CI is nearly twice the top of the IFHS CI. The factor-of-4 difference translates into 450,000 additional deaths in the L2 estimate above the IFHS estimate.

Even this formulation understates the difference between the two surveys. Using conventional estimation methods the IFHS estimate for violent deaths would have been below 100,000. The IFHS paper argues that conflict mortality surveys tend underestimate violent deaths and adjusts its conventional estimate up to 151,000. If this is right then, for a proper comparison, either the L2 estimate should be adjusted up similarly to how the IFHS estimate was adjusted up or we should compare unadjusted IFHS figures with unadjusted L2 figures. Making the latter comparison suggests at least a factor-of-six difference between L2 and the IFHS. Indeed, L2 estimated a violent mortality rate of 7.2 per 1,000 per year compared to a rate of 1.09 in the IFHS. These two estimates differ by a factor of 6.6. This translates into an L2 estimate that exceeds an unadjusted IFHS estimate by well over half a million violent deaths.

It is clear from much of the discussion above that the IFHS and the ILCS had more rigorous quality control than L2 did. Both the IFHS and the ILCS are also much larger surveys than L2. The IFHS interviewed 9,345 households in 971 clusters and the ILCS interviewed 21,668 households in 2,200 clusters compared to (as actually used) 1,849 households in 47 clusters for L2. In short, the ILCS and the IFHS are bigger and higher-quality surveys and both suggest that L2 has overestimated violent deaths by a wide margin.

[26] The IFHS recorded deaths occurring as late as June 30, 2006. L2 had a single cluster that recorded deaths occurring in July of 2006, L2's cluster 33 which is discussed in sub-section 3.5, but otherwise only covered through June of 2006.

I now compare the geographical patterns of deaths in the ILCS and L2. Table 1 shows that L2 and the ILCS agree rather well on violent deaths in the North and in the South.[27] In Baghdad L2 looks rather high compared to the ILCS but not exceptionally high. But in the central governorates L2 is very high indeed. Even when we allow only L2 deaths occurring before April of 2004, L2 still exceeds the upper limit of the ILCS CI by more than a factor of 7. This becomes a factor of 23 in Diyala governorate. .

Table 1. Violent Deaths: ILCS vs. L2 - March, 2004

	ILCS lower CI limit	ILCS central estimate	ILCS upper CI limit	L2 central through March 31, 2004	(L2 central)/ (ILCS upper limit)
Total	18,000	23,500	**29,000**	**68,000**	**2.3**
North	0	500	1000	0	0
South	8,000	12,000	16,000	13,000	0.8
Baghdad	4,000	7,500	11,000	14,000	1.3
Center	2,000	3,500	**5,500**	**41,500**	**7.5**
Nineveh	0	500	**1,000**	**3,500**	**3.5**
Al-Tameem	0	0	500	0	0
Diala	0	500	**1,000**	**23,000**	**23.0**
Al-Anbar	500	2000	**3000**	**8,500**	**2.80**
Salahuddin	0	1000	**1500**	**6500**	**4.30**

[27] The North includes Suleimaniya, Erbil and Dohouk and the South includes Babil, Kerbala, Al-Najaf, Al-Qadisiyah, Thi-Qar. Missan, Basrah and Al-Muthana.

Table 2 shows how much more L2 diverges from the ILCS when we extend L2 through to the end of May, 2004.

Table 2. Violent Deaths: ILCS vs. L2 - May, 2004

	ILCS lower CI limit	ILCS central estimate	ILCS upper CI limit	L2 central through May 31, 2004	(L2 central)/ (ILCS upper limit)
Total	18,000	23,500	**29,000**	**89,000**	**3.1**
North	0	500	1000	0	0
South	8,000	12,000	16,000	13,000	0.8
Baghdad	4,000	7,500	11,000	15,500	1.4
Center	2,000	3,500	**5,500**	**60,500**	**11.0**
Nineveh	0	500	**1,000**	**5,500**	**5.5**
Al-Tameem	0	0	**500**	**3,500**	**7.0**
Diala	0	500	**1,000**	**27,000**	**27.0**
Al-Anbar	500	2,000	**3,000**	**18,000**	**6.0**
Salahuddin	0	1,000	**1,500**	**6,500**	**4.3**

To summarize the patterns:

1. Nonviolent deaths match up well, ILCS versus L2.

2. Violent deaths also match up well between the two surveys in the North and in the South.

3. In Baghdad L2 is definitely high for violent deaths but not dramatically out of line with the ILCS.

4. In the center L2 has far more violent deaths than the ILCS.

The ILCS seems to perform perfectly well relative to L2 in discovering non-violent deaths throughout Iraq. The ILCS also seems to be just as capable as L2 in discovering violent deaths in the North and South. Therefore, we cannot argue that the ILCS, perhaps due to weaknesses in its questionnaire, was not as good as L2 in finding deaths that have truly occurred. The discrepancy only arises for violent deaths in one particular region where the sudden large distance of L2 from the ILCS casts doubt on L2.

This surplus of violent deaths in a single region should be viewed within the context of the refusal of the L2 authors to release data tying households to anonymized interviewer IDs. It is possible that a single interview team did all or many of the clusters into which so many of L2's violent deaths are packed.

The IFHS-L2 comparison also seems to confirm the L2 pattern of the lumping of deaths into the central governorates, although data are not yet available to repeat the precise L2-ILCS comparisons presented above. Figure one of the IFHS paper shows that L2 places about 26% of its violent deaths in Baghdad compared to 54% for the IFHS. About 65% of L2's deaths are in governorates in the center and south (Al-Anbar, Diyala, Nineveh, Salahuddin, Babylon and Basra), according to the classifications of the above tables, compared to about 35% for the IFHS.

Figure 1 of the IFHS paper also shows that the geographical pattern of deaths in the IBC database, which is based primarily on monitoring of the international media, is consistent with that of the IFHS but not with L2.

The IFHS paper also compares its estimates with L2's for three different time periods. The ratio of violent mortality rates for the two studies is 1.8 (not statistically different from 1) for March 2003-April 2004, 4.2 (highly significant) for May 2004-May 2005 and 7.2 (highly significant) for June 2005-June 2006. In short, L2 exhibits an extremely sharp upward trend over time compared to the relatively flat trend exhibited by the IFHS.[28]

Both the geographical and the temporal heaping of deaths in L2 are consistent with a hypothesis of fabricated/falsified data. The large divergence of L2 from the IFHS comes after the time periods covered by the two main surveys that existed when L2 was published: L1 and the ILCS. If falsified violent deaths were added into the L2 dataset it would make sense to add most of them after the time period for which comparisons with other surveys were possible at the time L2 was published. This could explain why L2 diverges from the IFHS much more strongly after the ILCS/L1 period than it does before.

L2's geographical departures from the ILCS and the IFHS come in governorates that are known to be violent but that are outside of Baghdad. L2 researchers knew that their estimates would be compared to the counts of the IBC's. A case can be made that the international media, the main source for IBC, covers Baghdad better than it covers other parts of the country. This may or may not be true but it is a claim that certainly sounds plausible.[29] If we accept the idea of Baghdad bias in IBC data then adding many falsified violent deaths into Baghdad clusters of L2 would create a very large L2/IBC divergence in Baghdad which would have been flagged as suspicious. Adding falsified deaths into zones known to be peaceful, such as the Kurdish area, would have also raised suspicions. A better strategy would be to add falsified deaths into acknowledged violent areas outside

[28] The fairly flat trend of the IFHS is relatively consistent with the daily data of the IBC, although IBC increases somewhat more sharply than the IFHS does in the final 13-month period compared to the second 13-month period http. The big upsurge in killing after the bombing of the Golden Mosque began in February of 2006, i.e., just before the end of the IFHS and L2 surveys, too late to produce L2's very sharp trend up over the last to 13-month periods.

[29] According to Burnham et al. (2006b) "Much violence is occurring far from the view of journalists and widely cited mechanisms for counting the dead. Most Western reporters are based in Baghdad." This comment overlooks the point that IBC includes many non-Western sources, often as translated by the BBC but still will resonate with many readers.

of Baghdad. i.e., the central governorates of Al-Anbar, Diyala, Nineveh and Salahuddin where L2 is so far out of line with the other data sources. The geographical pattern of deaths in L2 is, therefore, consistent with a falsification hypothesis.

Finally, note that the L2 paper claims that L1 and L2 confirm each other but Gourley et al. (2007) documents that this claim does not withstand scrutiny. The L2 data suggests roughly twice as many violent deaths during the L1 coverage period than were estimated in L1.

3.5. Cluster 33

The following anomaly was discovered by Olivier Degomme and Deberati Guha-Sapir of CRED in Belgium. They found that 24 people were killed by car bombs in July of 2006 in a single cluster of the L2 dataset: cluster 33 in Baghdad.[30] L2 field work finished on July 10, 2006. Therefore, these deaths must have occurred between July 1, 2006 and July 10, 2006. During this time period IBC recorded separate car bombings in which the number of people killed were 68, 17-19, 10-12, 6, 5 and fewer scattered through the neighborhoods of Sadr City, Adhamiya, Jameela, Mansour and Al-Bayaa respectively plus other places around Baghdad. It is crucial to note that, according to the L2 methodology, in each cluster a field team did interviews in 40 *contiguous* households. It is, therefore, exceptionally implausible that so many close neighbors could have been killed in multiple car bombings in different neighborhoods of Baghdad within a single 10-day window.[31] Thus, the most favorable interpretation for L2 is that all 24 victims were killed in the very large car bombing in Sadr City (BBC, 2006) and so I will assume this.

The pictures at BBC (2006) show rather clearly that there was not a line of homes destroyed.[32] It would seem to be virtually impossible for a group of 24 people coming from 18 separate homes located more or less right next to each other to all have been walking around the market clustered so close to one another when the bomb exploded. It is hard to imagine how this could have happened unless this large group of people all set out together for the market and then circulated through the market doing their shopping while holding hands. It seems likely that all or most of these deaths in the L2 dataset are fabricated.

Recall the evidence already presented on security-caused failures to visit clusters, L2 versus IFHS. I argued that the L2 claim of 12 successful Baghdad visits in 12 attempts was highly unlikely given the 67.7% success rate in cluster visits of the IFHS in

[30] These deaths were neatly arranged across households; 12 households had 1 death and 6 households had 2 deaths, a fact that is a bit suspicious in its own right.
[31] In fact, even the possibility of multiple neighbors killed in multiple car bombings in a single neighborhood is exceptionally implausible.
[32] It is very unlikely, but perhaps not impossible, that the international media, and hence IBC, might have overlooked some lethal car bombs in Baghdad. However, for the cluster 33 data to become plausible the international media would have to have missed a large car bomb that seriously damaged at least 18 homes while killing 2 inhabitants of 6 of them and 1 inhabitant of 12 of them.

Baghdad. Cluster 33 adds a specifically suspicious cluster to the general cloud that hangs over all of L2's Baghdad clusters in light of the IFHS.

It is important to see the anonymized interviewer IDs for all the clusters in L2 and to check the extent to which the same interviewers might have been involved in both cluster 33 as well as in other suspicious clusters, particularly in the governorates of Diyala, Al-Tameem, Al-Anbar, Nineveh and Salahuddin. Unfortunately, the L2 authors continue to withhold these data.

3.6. Death certificates

The very high rates of violent deaths measured in L2 have been defended on the grounds that a high percentage of the deaths recorded by L2 were confirmed through death certificates. According to the L2 paper and Burnham (2007):

1. Field teams requested death certificates for 545 out of 629 (87%) of deaths.

2. When field teams did not request death certificates this was because they "forgot" (Burnham, 2007).

3. When requested, respondents produced death certificates 501 out of 545 times.

4. "The pattern of deaths in households without death certificates was no different from those with certificates." (Burnham et al., 2006a)

The claim that a very high percentage of the deaths in the sample were confirmed by death certificates has been central to the defense of L2 from the beginning. Given the strong unpopularity of the occupation of Iraq it is easy to imagine that many respondents might have invented deaths.[33] Less dramatically, it seems likely that people might have reported deaths of extended family members who did not reside within the households of respondents. Very few respondents, and perhaps not even all of the interviewers themselves, would understand the statistical imperative to clearly limit household boundaries. To the contrary, many people will feel a need to "bear witness" to atrocities that have been visited on their friends and relatives. Many people may believe that the correct and moral thing to do is to report deaths of friends and family members. Such people might be baffled by the concept that somehow it is improper to report the death of, for example, a dear cousin.

L2 largely pre-empted such lines of criticism by claiming that their teams requested death certificates for 545 out of 629 (87%) deaths and respondents were able to produce them in 501 out of these 545 cases (92%).

[33] Recall that LSHTM (undated) advises that L2's approach of simply asking respondents how many household member they have and how many have died, rather than fully enumerating all household members with ages and genders, invites respondents to give "intentionally distorted responses".

There are, however, some reasons to question the high rate of death-certificate confirmation reported in L2.

1. The very high number of estimated deaths in L2 implies that the official death certificate system has issued, but failed to record the issuance of, about 500,000 death certificates during the L2 coverage period.[34] This forces L2 into a very delicate balancing act. For the death-certificate data to be valid it must be the case that Iraqi authorities issue death certificates for virtually all violent death and yet that same system fails to record the fact that death certificates have been issued roughly 90% of the time. Alternatively, it could be that Iraqi Ministry of Health is engaged in a massive and highly successful cover-up of deaths that have actually been documented through death certificates. This seems unlikely.

2. L2 had an extremely compressed work schedule. Field teams routinely had to complete 40 interviews in a day. This means that respondents had to produce these death certificates almost without fail and within a matter of minutes. In many cases these documents would not have been accessed for several years prior to an L2 interview.

3. In L1, the previous *Lancet* publication on Iraq by (mostly) the same team, the claimed rate of death certificate confirmation upon request was substantially lower than in L2: 80% when requested in L1 compared to 92% when requested in L2. The coverage period for L2 is nearly two years longer than the recall period for L1 so it should have been, if anything, harder to confirm deaths through death certificates in L2 compared to L1. Moreover, a significant fraction of the population had migrated during the time between the two studies with, presumably, at least some death certificates mislaid or buried amongst other belongings during these movements.

With the release of some L2 data it became possible to examine L2's death-certificate claims further. Here are some relatively new findings on death certificates mixed with some older discoveries from Kane (2007).

In the table 3 below "no" means that a death certificate was requested but not produced, "yes" means that a death certificate was requested and produced and "forgot" (consistent with Gilbert Burnham's MIT lecture) means that a death certificate was not requested.

[34] See "Implication 4" of Dardagan et al. (2006b) and Roug and Smith (2006).

Table 3. Death-Certificate Confirmation and Non-Confirmation of Deaths in L2

Governorate	No Violent	No Non-Violent	Yes Violent	Yes Non-Violent	Forgot Violent	Forgot Non-Violent
Babil	**0**	**0**	6	22	**0**	**0**
Kerbala	**0**	**1**	3	5	**0**	**0**
Wasit	**0**	**0**	0	5	**0**	**0**
Al-Najaf	**0**	**2**	0	14	**0**	**0**
Al-Qadisiya	**0**	**0**	4	11	**0**	**0**
Thi-Qar	**0**	**11**	4	15	**0**	**0**
Missan	**0**	**0**	3	7	**0**	**0**
Basra	**0**	**1**	16	35	**0**	**1**
Suleimaniya	**0**	**2**	0	6	**0**	**0**
Erbil	**0**	**1**	3	18	**2**	**0**
Baghdad	**0**	**0**	27	73	**50**	**10**
Nineveh	**22**	**2**	30	34	**7**	**0**
Al-Tameem	**0**	**0**	0	1	**2**	**2**
Diala	**0**	**3**	51	18	**3**	**0**
Al-Anbar	**0**	**0**	38	19	**6**	**0**
Salahuddin	**0**	**0**	25	8	**0**	**0**

It is clear that, contrary to the claims of L2, the pattern of deaths with death certificates does differ from those without.

1. For violent deaths all failures to produce death certificates when asked were in a single governorate, Nineveh, whereas for non-violent deaths these failures were spread across eight governorates. It is implausible that the system of issuing death certificates and families taking care of them is nearly perfect in all but one governorate in the case of violent deaths whereas these systems are less reliable for non-violent deaths in 8 governorates.

2. "Forgetting" to ask, or simply not asking, was far more common in Baghdad than outside Baghdad and six times more likely overall for non-violent deaths than for violent deaths (Kane, 2007).

3. Baghdad, Nineveh and Thi-Qar all display strange patterns and need to be examined more closely.

Under a variety of reasonable assumptions the perfect run of 180 death certificate confirmations in 180 attempts for violent deaths outside Nineveh appears to be extremely unlikely, e.g.,:[35]

[35] I assume statistical independence across deaths for all of these calculations.

1. Using the death-certificate confirmation rate for L1 of 80% and assuming statistical independence across deaths, the odds against 180 confirmations in a row are $2.7x10^{27}$ to 1. In fact, a more direct comparison is possible for the violent deaths recorded in L2 and occurring during the L1 coverage period, i.e., through September of 2004. L2 claims a perfect record of 60 confirmations in 60 attempts for violent deaths during the L1 sampling period, for which we can calculate odds of more than 650,000 to 1 against.

2. Using the confirmation rate for non-violent deaths in L2 of 92%, the odds against are more than three million to 1.

3. Even if we arbitrarily and implausibly assume a 0.98 probability that death certificates can be produced for each violent death we still get odds of 38 to 1 against.

I conclude that there is likely fabrication in the death-certificate data in L2 and that these data do not give reliable support to L2's very high estimated death rate.

3.7. Cluster 34

As noted in Section 3.6, L2 reports that its respondents failed to produce death certificates when asked only 22 times for violent deaths. All 22 of the missing death certificates for violent deaths occurred in the governorate of Nineveh. L2 has 5 clusters in Nineveh. One of these, Cluster 34, contains 19 of these 22 confirmation failures.

Cluster 34 contains 42 deaths, 35 of which are classified as violent. These violent deaths break down into18 by "air strike", 10 from "gunshot", 4 from "car bombs", 1 from "fight", 1 from ""crushed, USA Army Vehicle" and 1 from "bomb".

The 18 deaths in air strikes, which could only be due to the USA, contribute about 36,000 deaths to L2's central estimate of 600,000 violent deaths. According to the L2 dataset none of these deaths were confirmed by a death certificate. For 7 of the 18 the interviewers forgot to, or simply did not, ask for death certificates. These 7 were in a single household that reported deaths of 2 girls, 3 boys and 2 women (one aged 17), due to an air strike taking the specific form of a "missile on home" in November of 2005.

For all of the remaining 11 deaths from air strikes in cluster 34 it is reported that interviewers asked to see deaths certificates but respondents were unable to produce any. These include a second household that reported deaths in November of 2005, 2 boys under the age of 5, possibly in the same event as the above "missile on home" that is claimed to have killed 7 women and children in the same month. The L2 dataset claims 4 further air strikes in cluster 34. These events were in June of 2005, killing 2 men in a single household; in October of 2005, again killing 2 men in a single household; in December of 2005, killing l girl; and in March of 2006, killing 2 men in one household and 2 girls in another household.[36]

[36] One of the victims of the October, 2005 air strike was a 15-year-old male, classified as an adult in L2.

Cluster 34's 18 deaths in air strikes are spread over 7 households in 5 different months. Thus, according to the L2 data there were at least five separate air strikes on this small neighborhood of 40 contiguous households over a 10-month period between June of 2005 and March of 2006. All of these air strikes came months after the first few weeks of the war in 2003 when air strikes were common.

Claimed air-strike victims in cluster 34 include 2 women and 10 children spread across 4 households in at least three incidents plus a 15-year-old in a fifth household/fourth incident. Survivors in all 5 of these households would have strong motives to report these deaths so as to receive financial compensation from the United States. Thus, if real, these deaths would be more likely to be backed by death certificates than most deaths in Iraq would be. Yet L2 reports that none of these deaths were corroborated by death certificates. It is also likely that 12 air-strike killings of women and children would draw international media attention. Yet none of these deaths appear in the IBC database, a strong indicator that they were not reported by the international media.[37]

Table 4. The Age Distribution of People Killed By US Air Strikes in Cluster 34

Age	2	3	5	7	9	13	14	15	17	19	22	41	49
Number Killed	2	1	3	1	1	1	1	1	1	2	1	1	2

Table 4 gives the age distribution of the victims of US air strikes in cluster 34. This is a surprisingly young set of victims, as many as 2/3 of whom could be considered children, with 3 of the remaining 6 aged 19 or 22. The complete absence of victims over the age of 50, or in their late twenties or thirties is puzzling. Of course, there exists a general and valid perception that it is worse to kill children than it is to kill adults. Thus, this age pattern is consistent with the hypothesis that respondents or interviewers fabricated deaths to make US soldiers look bad. Similarly, 1/3 of the claimed victims in these air strikes were female, although only 9% of all violent deaths in L2 were of females.

The 5 deaths attributed to "bullet by USA army" account for about 10,000 violent deaths in the L2 estimate. They break down into 2 adult males in separate households with death-certificate confirmation in February of 2005, a man in May of 2005, and a girl and a woman in single household in June of 2005. For the last three deaths it is reported that interviewers requested death certificates but respondents were unable to produce them. Unlike the claimed air-strike deaths, some weak corroborating evidence can be found for these shootings within the IBC database. IBC does have shootings involving US forces, sometimes in firefights with "anti-coalition agents," in the relevant months in various

[37] IBC records 8 deaths from an incident in Mosul on May 19, 2005, that included helicopter fire and could, therefore, be viewed as an air strike. Conceivably, this incident could match the June, 2005 incident coded in L2. Similarly, IBC has a September 5, 2005 air strike in Tal Afar, killing 6 and hitting several houses that could be stretched to match the cluster-34 incident of October of 2005. These 2 air strikes were in different cities so at most only one could match the claimed air strikes for cluster-34.

places within the governorate of Nineveh.[38] Nevertheless, it still seems unlikely that there were at least 3 separate shooting incidents in which US soldiers killed residents of 4 households in this small neighborhood of 40 contiguous households within a span of 17 months.

The final death attributed to the USA is a 3-year-old boy claimed to have been crushed by an American military vehicle in August of 2005 with death certificate confirmation. This death does not appear in the IBC database although it is a newsworthy incident if true.

There is no overlap between the 7 households reporting deaths from US air strikes, the 4 households reporting deaths from USA bullets, and the household reporting a child crushed by an American military vehicle. Thus, cluster 34 contains 12 households claiming 24 deaths attributed to the US military in at least 9 separate incidents over a 17-month period. These 24 deaths attributed to the US military in cluster 34 constitute fully one quarter of all violent deaths attributed to coalition forces in L2 and account for about 8% of all violent deaths in L2.

The 24 violent deaths at the hands of US soldiers are 69% of all the violent deaths in the cluster. In contrast, in the IBC database the US is coded as being fully or partially responsible for 476 out of 2,963 (16%) violent deaths of civilians in the governorate of Nineveh during the L2 sampling period. Cluster 34 contributed about 48,000 violent deaths blamed on US forces to L2's central estimate, roughly 100 times the number of civilian deaths fully or partially attributed to US forces by IBC in the entire governorate of Nineveh. But the true discrepancy is still larger since the L2 dataset contains 5 Nineveh clusters.[39]

The 24 people violently killed by US soldiers in Cluster 34 break down into 6 girls, 6 boys, 3 women and 9 men: 9 females and 15 males. Thus, in Cluster 34, 50% of these US victims were children and 38% were females. In contrast, of all violent deaths in the full L2 dataset, 11% were children and 9% were females. In all clusters combined 19 out of 95 US victims (20%) were children and 12 (13%) were females. The entire L2 dataset contains 50 violent deaths of women and children, 15 of which (30%) are recorded as killed by the USA in cluster 34 alone.[40] According to the L2 dataset, in cluster 34 alone the US military killed 3 of the 16 women (19%), 6 of the 22 boys (27%) and 6 of the 12 girls (50%) killed violently by any party in all of L2's 47 counted clusters combined. To summarize, if the cluster-34 data are true, the behavior of US soldiers within the cluster was much worse than the behavior throughout the whole of Iraq both of US soldiers themselves and of all other agents.

[38] Matching events by governorate within a time frame of one full month provides only weak corroboration.
[39] The central estimate of the IFHS for civilians and combatants in all of Iraq is roughly three times the IBC estimate for violent deaths of civilians. Extrapolating this factor of three to cover killings by the US in Nineveh would implies that L2 overestimated killings by US soldiers in Nineveh by much more than a factor of 30.
[40] L2 mistakenly reports "Of the 302 violent deaths, 274 (91%) were of men…" but the 274 violent deaths of males break down into 252 men and 14 boys.

A number of factors already presented are all suggestive of fabrication of violent deaths in cluster 34. These include: 1) the number of killings attributed to US soldiers in the cluster; 2) the number of incidents of such killings; 3) the unique focus of these killings on women and children, compared both to killings by other agents in Iraq and to US norms throughout the country and; 4) the thinness of corroborating evidence for these killings, either through death certificates or through the international media.

There is further evidence of fabrication in the fact that 19 out of the 24 deaths attributed to the Coalition in cluster 34 are claimed by a string of 9 households with L2 dataset IDs of 1311, 1312, 1313, 1314, 1315, 1317, 1319, 1320 and 1321. To the extent that consecutive numbers within the dataset suggests that households are in particularly close proximity with each other, this pattern suggests that there may have been some coordination among neighbors on reporting fabricated violent deaths caused by the US. Such coordination could have been facilitated by advance approaches by neighborhood children, as discussed in Section 2, to explain the purpose of the L2 survey. Alternatively, this string of households might have been interviewed by a single interview team that was fabricating deaths.

Cluster 34 contains an additional 11 deaths not directly attributed to the US. Of these, 5 come in bombings, 4 of which are specifically classified as car bombings. These deaths are spread over 4 new households, i.e., households not reporting deaths caused by the US, and three separate months. The first car-bomb killing was of a man in April of 2005 claimed to be verified by a death certificate. Next, in November of 2005 there were car-bombing deaths of 1 man and 1 woman. In both cases it is reported that death certificates were requested but not produced. Also, in November of 2005 there was a bombing death of a 15-year-old classified as a man. These November bombings may have been the same event although they victimized two separate households. The fifth death was a man from a fourth household in May of 2006 and again it is reported that a death certificate was requested but not produced. The international media did report multiple car bombings in Nineveh in April of 2005 and May of 2006 so there is some small corroboration, at least for 2 of the 3 car bombings.[41] Nevertheless, it is very unlikely that five people spread across four separate households within a small group of 40 adjacent households would have been killed in three separate car bombings. The probability of this happening may well be lower than the probability that 24 members of a single cluster could have been killed in a single car bombing, as is claimed for cluster 33.

L2 claims five further gunshot deaths, all of men, in cluster 34 in addition to the 5 people shot to death by US soldiers already discussed above. In March of 2004 there was a "gunshot robbery" of a man claimed as verified by death certificate. There were four subsequent deaths in the cluster from "gunshot unknown." The first 2, in November and

[41] As noted above, matching events by governorate within a time frame of a month is weak corroboration. Even such corroboration is not possible for the third car bombing. The IBC database contains car bombings in October and December of 2005 but none in November of 2005, when L2 claims 3 bombing deaths in cluster 34. Of course, it is possible that some car bombings are missed by the international media and/or by IBC. However, car-bombings are highly visible and newsworthy and both insurgents and coalition forces have strong incentives to report them. Therefore, it is unlikely that very many, if any, lethal car bombings are overlooked.

December of 2004, are coded as verified by death certificates. For the second 2, in September of 2005 and April of 2006, it is reported that death certificates were requested but not produced. None of these overlap with any of the above incidents or households. Thus, they yield five further incidents affecting five further households among this small cluster of 40 contiguous households. IBC has a number of gunshot deaths attributed to "anti-coalition agents" and "unknown agents" during each of these months. Nevertheless, so much targeting of this one small neighborhood seems unlikely. Remember, that the L2 authors claim, in various forms, that all neighborhoods had essentially equal chances of being selected into the sample.

The final violent death in cluster 34 was a man from another new household recorded as dying in a "fight" confirmed by a death certificate in November of 2004. Conceivably this was the same incident in which a member of a different household died from a gunshot.

The 11 violent killings not directly attributed to US soldiers in cluster 34 break down into 10 men and 1 woman, although one man was only 15 years old. Thus, the percentage of females killed among these 11 deaths, 9%, exactly matches of the percent of females killed among all violent deaths in the L2 dataset.[42] Table 5 summarizes how the number of violent killings plus their gender and age mix compare for US soldiers and for other agents both within cluster 34 and for all clusters. If true, it points to exceptionally dirty behavior for US soldiers in cluster 34 where the US is blamed for about 1/2 of all killings of women and children nationwide by L2. Other agents are held responsible for killing 1 woman and no children.

Table 5. People, Females and Children Killed by US Soldiers and Other Agents

	Killed in Cluster 34	% Killed in Cluster 34	% Killed in all Clusters	Children Killed in Cluster 34	% Children among all Children Killed in all Clusters	Females killed in Cluster 34	% Females among all Females Killed in all Clusters	Girls killed in Cluster 34	% Girls among all Girls Killed in all Clusters
US Soldiers	24	69%	31%	12	46%	9	32%	6	50%
Other agents	11	31%	69%	0	0%	1	3.6%	0	0%

Combining the violent activity of US soldiers and other agents, cluster 34 contains at least 17 separate violent incidents affecting 22 of the 40 households in the cluster and causing 35 violent deaths. It is reported that only 9 of the violent deaths were confirmed by death certificates, i.e., about 26%. Of the 26 non-corroborated violent deaths, death certificated were not requested for 7 (27%) and were requested but not produced for 19 (73%).

[42] Obviously, the percent of children violently killed, 0%, is below the average of 11% for the L2 dataset as a whole. However, this figure is based on small numbers and would more or less reach the average if the 15-year-old were reclassified as a child.

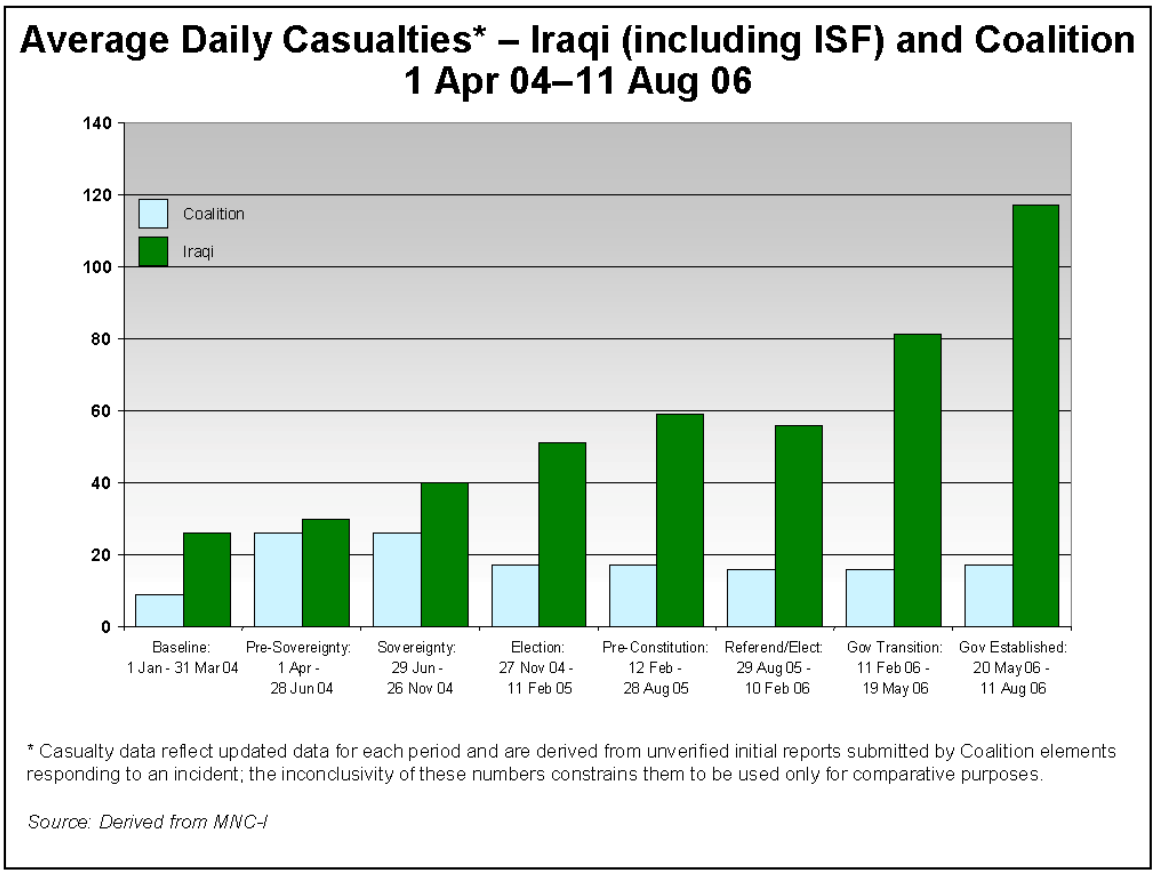
Evidence of fabrication of violent deaths in this small cluster of 40 contiguous households comes in four basic forms. First, cluster 34 seems to have been afflicted with improbably large numbers of violent deaths, violent incidents and households affected by this violence. Second, the extent to which and manner in which US soldiers are blamed for these killings suggests some attempts to tarnish the reputation of US soldiers. The total numbers of US victims, female victims and child victims in cluster 34 are large compared to the victims of other agents in the cluster. The percentages of female and child victims of US soldiers among all female and child victims of all agents within cluster 34 are very high: 90% and 100% respectively. The percentages of female and child victims of US soldiers within cluster 34 among all female and child victims of all agents in all clusters are also very high: 32% and 46% respectively. For these claims to be true, the behavior of US soldiers in Nineveh would have to be very much worse than the behavior of other agents in Nineveh and normal behavior of US soldiers elsewhere. Third, there is no corroborating evidence, either through the international media or through death certificates, for many of the deaths. Fourth, there is a string of household ID's within which 9 households out of 11 reported killings by US soldiers, suggesting that there might have been a coordinated attempt, either by interviewers or respondents, to manipulate the L2 survey.

3.8. Mishandling of other evidence on mortality in Iraq

The L2 paper ignores contrary evidence, creates spurious confirming evidence and manipulates other evidence on mortality in Iraq. The impact of these distortions is to obfuscate the extent to which L2 is an outlier among all the credible sources of mortality information in Iraq (see also Spagat, 2008).

The L2 introduction contains at least the following problems:

1. It cites the US Department of Defense (DoD) as recording 117 civilian deaths per day between May, 2005 and June, 2006. But, Dougherty (2007) exposed the fact that, the source cited, DoD (2006), states clearly that this figure is 117 casualties per day of civilians plus combatants (Iraqi Security Forces) where casualties means killings plus injuries. The original figure from the DoD report is reproduced below. Note also that the DoD figure of 117 actually applies to the period May 20, 2006 through August 11, 2006, not May, 2005 through June 2006 as claimed in L2. To cover the period of May, 2005 through June 2006 cited in L2 we need to include three other periods during which casualties per day of Iraqi civilians plus combatants are, respectively, roughly 82, 55 and 59. Thus, the DoD figures suggest perhaps 70 casualties per day of civilians plus combatants during the period cited in L2, a difference of more than 20,000 casualties. Civilian deaths measured by DoD are likely to be considerably lower than 117 per day during the appropriate period. This is, in fact, a period when L2 measures roughly 1,000 violent deaths per day. The DdD figures are again misrepresented as mortality numbers in figure 4, later in the L2 paper.



2. It ignores the fact that the ILCS estimated war-related deaths and that its figures are much lower than the L2 figures. As noted above, the L2 estimate exceeds the ILCS one by a factor of 3 or 4. L2 mentions the ILCS but only as confirming that bad water, sewerage and restricted electricity create health problems. L2 also mentions the ILCS in

a footnote as "predictably" finding substantially higher numbers than what L2 refers to as "passive surveillance" efforts, i.e., IBC. Yet the ILCS estimate for civilians plus combatants killed is only 1.6 times the IBC number for only civilians killed during the ILCS period. This period is the early phase of the war when many combatants were killed. L2, on the other hand, differs by a factor of 12 with IBC, somewhat less if we take some account of combatants.

3. It ignores the UN mortality monitoring (UNAMI, 2007). These figures are lower than the L2 figures for 2006 by about a factor of 12 during the first half of 2006. UNAMI measured about 80 deaths per day compared to about 1,000 per day for L2 or about 170,000 violent deaths in L2 supposedly missed by the UN monitoring system.

4. It ignores the daily casualty monitoring of the Iraq Ministry of Health Emergency Tracking System (Sloboda et al., 2007). These figures are lower than L2's by about a factor of 15.

5. It does mention the IBC figures, which are lower than L2's by a factor of 12, but does not compare them to L2. Instead, L2 gives a misleading comparison suggesting that the figures of Iraq's Interior Ministry are 75% higher than IBC's, which might suggest to some readers that the IBC figures should be dismissed as far too low:

> "Estimates from the Iraqi Ministry of the Interior were 75% higher than those based on the Iraq Body Count from the same period." (Burnham et al., 2006a)

In fact, IBC figures are 50% higher than the Interior Ministry figures to which they are compared in the cited source (O'Hanlon and Kamons, 2006). On close inspection we see that this is an effort of the Brookings Institution that removes all morgue entries and police deaths from IBC. These figures are then compared in L2 to Interior Ministry figures which would likely include police and morgue data. This manipulation brings the IBC figures from 50% above to 40% below the Interior Ministry ones.

6. It cites L1 as confirming L2 but, as noted above (Gourley et al., 2007), this is not the case.

7. It comments that in many conflicts indirect and nonviolent deaths comprise the majority of excess deaths. Yet it fails to mention that L2's findings conflict with this common pattern. Excess non-violent deaths are statistically insignificant in L2.

8. It cites (Janabi, 2006) claiming that "a detailed survey" had been conducted that found 37,000 civilian fatalities between March 2003 and September 2003 in Iraq. The origin of this *Aljazeera* story was a letter posted on a blog on August 21, 2003 (Wanniski, 2003) claiming that the Iraqi Freedom Party had made a massive census-like effort to collect data on civilian deaths, visiting

> "all villages, towns, cities and some of the desert areas etc. affected by the aggression (with exception of the Kurdish area), and also by interviewing hundreds of undertakers, hospitals officials and ordinary people in these places, conducted a survey (Wanniski, 2003)

The posting goes on to explain that the sole copy of the report on this survey was in the possession of a single man who was unable to find a fax machine (or apparently a photocopying machine) in Baghdad so that he could fax the report to party headquarters. He had, therefore, attempted to cross over to the Kurdish zone of Iraq in search of a fax machine and had disappeared with the only copy of report. Apparently, all supporting materials from this massive effort are also lost so there will never be a new write-up:

> "Due to the absence in Iraq (with the exception of the Kurdish area) of functional communication systems with the outside World, our party headquarters in Baghdad tried to send me a fully comprehensive and detailed report by fax from Al-Sulaymaniyah (a Kurdish area). However, by crossing to the Kurdish area, the Kurdish "Peshmarga" searched the person carrying that report which was found with him and confiscated. According, he was handed over to the American troops where he was arrested and no one knows yet of his whereabouts." (Wanniski, 2003)

Such evidence is not suitable for citation as a credible source in an academic paper.

9. It claims, similarly, that

> "Iraqiyun, estimated 128,000 deaths from the time of the invasion until July 2005, by use of various sources, including household interviews". (Burnham et al., 2006a)

Yet in Appendix C of Burnham et al. (2006b) the L2 authors are less confident about this source:

> "The methods of this organization - reported to be direct accounts from relatives of those killed - could not be confirmed" (Burnham et al. 2006b)

Burnham et al. (2006b) cites UPI (2005):

> "An Iraqi humanitarian organization is reporting that 128,000 Iraqis have been killed since the U.S. invasion began in March 2003.
>
> Mafkarat al-Islam reported that chairman of the 'Iraqiyun humanitarian organization in Baghdad, Dr. Hatim al-'Alwani, said that the toll includes everyone who has been killed since that time, adding that 55 percent of those killed have been women and children aged 12 and under." (UPI, 2005)

This three-paragraph UPI article is the sole basis for the claim that a survey was done. No copy of the survey has every surfaced. Cole (2007) refers to Mafkarat al-Islam as "The radical Sunni Arab newspaper." This is what the US State Department has to say about Mafkarat al-Islam (Islam Memo):

> "*Islam Memo*, or Mafkarat al-Islam, is perhaps the most unreliable source of "news" about Iraq on the Internet. For example, on March 27, 2005, *Islam Memo* "news items" translated into English by Muhammad Abu Nasr claimed that more than 88 U.S. soldiers had been killed that day. In reality, none had been killed. Such disinformation fabrications are typical of *Islam Memo*. In the ten-day period from March 20 to March 29, 2005, they claimed that more than 334 U.S. troops had been killed. The real number was eight." State Department (2005).

L2 diverts readers from this trail by not citing the UPI article but instead citing NGO Coordination Committee of Iraq (2006), a fourth-hand reference which gives the Iraqiyun figure, citing the *Washington Times* which, in turn, just reprinted the UPI article.[43]

These problems all arise just within the first 4 paragraphs of the L2 paper. They show a consistent pattern of ignoring or misconstruing contrary evidence, claiming supporting evidence that is not fit for citation and creating supporting evidence from sources that do not actually support L2. These practices follow a pattern already established in Checchi and Roberts (2005) which created a misleading table (table 6) that conveys a false impression that analysis of seven selected mortality sources for Iraq showed that IBC's figures were low by factors of five to ten and those of L1 were moderate. Among other problems, this table cuts the IBC numbers almost in half and cites a mental health study published in the *New England Journal of Medicine* as yielding an extremely high mortality rate although the study offers no mortality estimate and its data are not usable for such a purpose. This effort was refuted in detail in Dardagan et al. (2006a). These are examples of information falsification.

L2's figure 4 attempts to convince readers that L2's extremely sharp upward trend in mortality rates from the beginning of the war until the middle of 2006 is consistent with evidence from both the DoD and IBC. It is claimed that these common trends support the credibility of the L2 data. Figure 4 is, however, incorrect and misleading. First, as noted above, the DoD figures are for casualties and not mortality so they are not comparable to the L2 ones. Second, the DoD figures only begin January 1, 2004 yet figure 4 claims a DoD figure of roughly 12,000 deaths covering March 2003 through April 2004. This figure of 12,000, which is placed virtually on top of the IBC figure, seems to be without any basis. Third, as pointed out in Guha-Sapir et al. (2007) figure 4 compares L2 numbers for deaths per 1,000 per year over three time periods since the start of the war with cumulative DoD and IBC figures. Of course, cumulative figures increase sharply, much like the L2 rates do. But a proper comparison of rates versus rates shows the IBC figures to be relatively flat over time while the L2 ones increase very sharply. The DoD casualty rates for the 13-month period June 1, 2005 through June 30, 2006 are about 45% higher than DoD figures for May 1, 2004 through May 31, 2005: 0.96 and 0.66 casualties per 1,000 per year respectively. The corresponding figures for L2, quoted in figure 4, over the same time periods are 10.9 deaths per 1,000 per year and 19.8 deaths per 1,000 per year, an 82% increase. Therefore, deaths in L2 increase more sharply than casualties in the DoD data. Yet, figure 4 places the DoD point below the L2 point for May 2004 through May 2005 and above the L2 point for June 2005 through June 2006, creating a false impression that the DoD data exhibit a sharper upward trend than the L2

[43] The sources run from *Islam Memo* to UPI to the *Washington Times* to the National Coordination Committee of Iraq to the *Lancet*.

Figure 4 as Printed in L2 Together with the Corrected Version

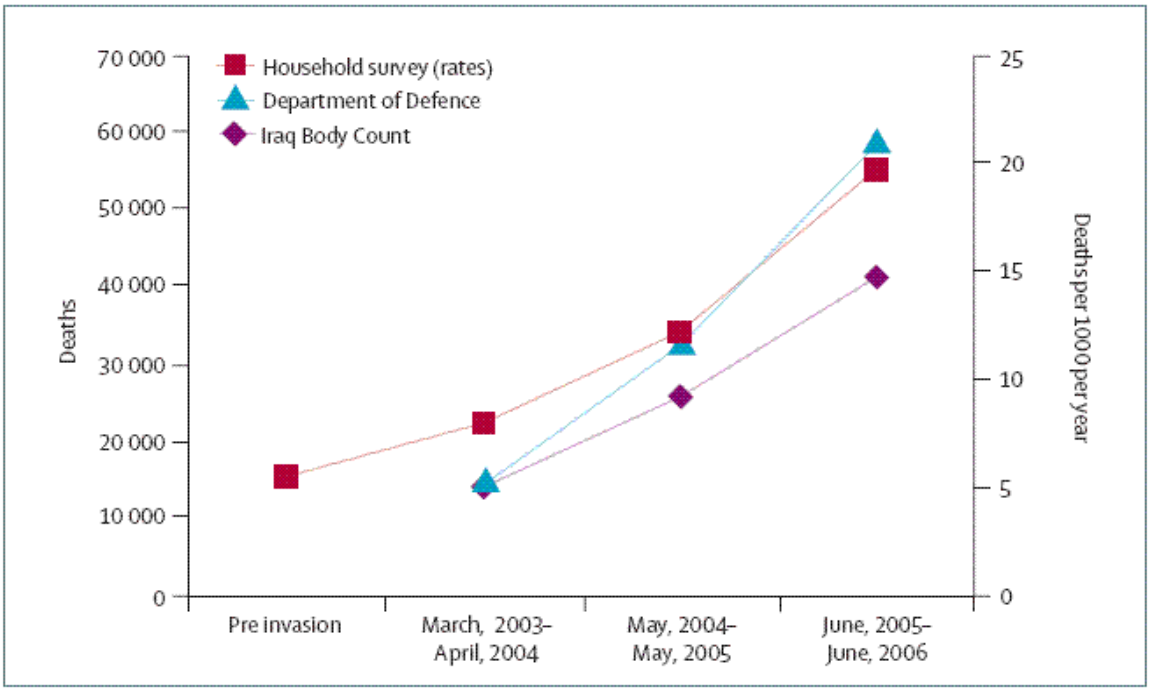
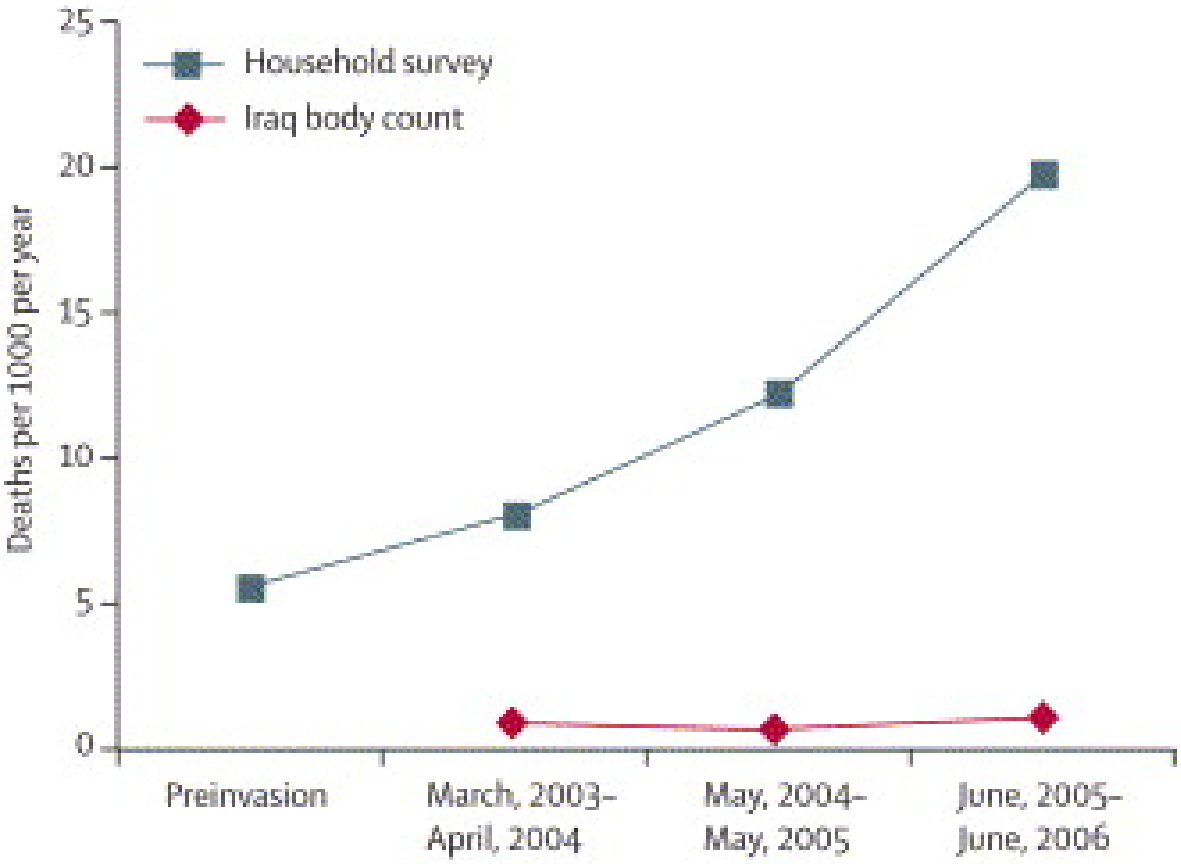


Figure 4: Trends in number of deaths reported by the Iraq Body Count and the MultiNational Corps-Iraq and the mortality rates found by this study



data do. The opposite is true. The previous page reproduces figure 4 as it appears in L2 together with the corrected figure.

Recall that I argued above that the very sharp upward trend for violent mortality rates in L2 after the L1 and ILCS sampling periods were finished is, in itself, suggestive of data fabrication. Figure 4 hides this anomaly.

There is further mishandling of evidence in the "Discussion" section of L2. The objective is to explain the huge difference between IBC figures (and also the spuriously cited DoD figures) and L2 figures by claiming that IBC's "passive surveillance"[44] methods have been shown to only capture a tiny fraction of all conflict violence:

> "Our estimate of excess deaths is far higher than those reported in Iraq through passive surveillance measures. [footnote to IBC and the DoD] This discrepancy is not unexpected. Data from passive surveillance are rarely complete, even in stable circumstances, and are even less complete during conflict, when access is restricted and fatal events could be intentionally hidden. Aside from Bosnia [footnote], we can find no conflict situation where passive surveillance recorded more than 20% of the deaths measured by population-based methods. In several outbreaks, disease and death recorded by facility-based methods underestimated events by a factor of ten or more when compared with population-based estimates. [five footnotes]. Between 1960 and 1990, newspaper accounts of political deaths in Guatemala correctly reported over 50% of deaths in years of low violence but less than 5% in years of highest violence. [footnote]" (Burnham et al., 2006a).

What are these allegedly supporting footnotes?

1. The "Bosnia" study cited is actually a Croatia study (Kuzman et al., 1993). The paper examines 4,339 deaths "recorded on two documents: a demographic mortality statistical form completed by authorized civil servants, and a death certificate completed by medical examiners." The paper cites Ministry of Health figures that estimate "a total war toll of 10,000 to 12,000 deaths or more" but does not say how the Ministry of Health made these estimates. It also mentions that the Red Cross counted 13,708 missing persons but does not speculate on how many of these people died. Conceivably, this paper could have some implications for official surveillance systems but it has no implications for media-based monitoring in Iraq.

2. Roberts et al., 2001, a study done in the DRC. It reports on a population-based survey but contains no comparison with any other figures derived from other methods. On its own, it cannot be used to argue that any method undercounts war deaths by any factor compared to population-based methods.

---

[44] The term "passive surveillance" seems to have originated in the medical literature to refer to data on medical ailments compiled by recording the number of people who present themselves to medical facilities for treatment. This is contrasted to "active surveillance" methods by which data collectors proactively search the community and find ailing people. Applying the "passive surveillance" term to conflict journalism is misleading since journalists actively seek out violent events, witnesses and informed sources in the field. It also does not apply well to DOD data collected by reports of soldiers on incidents in which they have engaged.

3. Roberts and Despines (1999), a letter on mortality in the DRC that reports only on survey findings and does not compare with any other figures.

4. Goma Epidemiology Group (1995), a study of the health of Rwandan refugees in Zaire. The study includes a survey but it is not used to estimate deaths. Thus, the paper makes no comparison of population-based estimates with deaths estimates from "passive surveillance". This work contains nothing that could be used to evaluate the coverage rate of media-based monitoring such as IBC's.

The paper seems to have been included as a supporting footnote because it does refer to undercounting of deaths:

> "48,347 bodies were collected by the trucks between July 14 and Aug. 14. This figure represents a minimum estimate for mortality in this population because an unknown, though probably small, number of refuges who died during the first few weeks were buried privately and, therefore, were not counted by the body collection system." (Goma Epidemiology Group, 1995)

The paper also observes that the area consists of hard volcanic rock so burial is difficult and bodies are normally left on the ground and are, therefore, easy to count. So this undercount, irrelevant for Iraq, is thought to be small in any case.

5. A study of a pellagra outbreak among refugees in Malawi in 1990 (Malfait, et al., 1993). Pellagra is a nutritional disease that can result in death in severe cases. This study is not relevant to mortality monitoring in Iraq. Violent killings in Iraq are an international news story. A normally non-fatal nutritional disease among refugees in Malawi is not an international news story. Coverage rates in the monitoring of pellagra in Malawi in 1990 cannot convey useful information about coverage of mortality monitoring in Iraq. In any case, although the article does discuss passive and active surveillance there is no direct comparison between the two since the two systems were never operated simultaneously.

6. Spiegel and Salama (2000), a population-based study of Kosovo that estimated 12,000 deaths. The study makes no mention of passive surveillance or media monitoring. It does mention three other estimates that range between 9,269 and 11,334, i.e., 77% to 94% of the study's estimate.

7. Ball et al. (1999), a Guatemala study, which is the only one mentioned that actually does compare some form of media monitoring with another method. Yet this analysis has little or no applicability to the IBC's mortality monitoring in Iraq. The Guatemala study argues that 13 mainstream newspapers in Guatemala failed completely to cover large massacres in the Guatemalan countryside in the late 1970's and early 1980's. On the other hand, it also notes that the international media and even some non-mainstream Guatemalan sources did convey at least some news about this violence. Although it is interesting to learn what the mainstream newspapers reported in Guatemala, this base of newspapers is too narrow to illuminate IBC's coverage of Iraq. IBC incorporates news wires, many non-mainstream news sources and official figures like those of the Baghdad morgue and the Ministry of Health. Moreover, Iraq at this point in time is far more in the

media spotlight than Guatemala was in the late 1970's and early 1980's and modern technologies like the internet and cell phones carry information much more freely out of Iraq in the 21st century than was the case in Guatemala nearly 30 years ago. Moreover, the killings in Guatemala during the relevant period were mostly of indigenous peoples' who were probably not prioritized by mainstream Guatemalan newspapers. Finally, according to the Guatemala study, mainstream newspapers captured more violence than the population-based measurements in a number of years. Thus, the Guatemala study does not imply that we should expect a coverage rate for IBC on the order of 5% as suggested in L2.

The following comparisons are not included among these L2 footnotes despite being far more relevant to the case of Iraq than the articles cited. They all suggest substantially more than 20% coverage for media-based monitoring in Iraq, contrary to the L2 claim that "we can find no conflict situation where passive surveillance recorded more than 20% of the deaths measured by population-based methods":

1. L1, conducted by mostly the same authors as L2, estimated 56,700 violent deaths of civilians plus combatants outside Al-Anbar governorate (EPIC, 2004), a large outlier in L1, compared to 17,687 deaths of civilians in Iraq outside Anbar recorded by IBC for the L1 period.

2. The ILCS estimated 24,000 war-related deaths of civilians and combatants compared to an IBC figure of about 14,000 deaths of civilians for the ILCS coverage period.[45]

3. Benini and Moulton (2004), a study of Afghanistan done by colleagues of the L2 authors at Johns Hopkins, compared mortality estimates from a population-based survey with a body count based on media monitoring that used methods that inspired IBC's approach (Herold, 2004). The survey found 5,576 killed. This compares to a media-based count of 3,620 civilians killed for the same period.

I draw two conclusions from the material discussed in this section. First, L2 is much more of an outlier in the Iraq mortality literature than would be suggested by L2's treatment of the literature. Second, the treatment of the evidence on Iraq mortality in L2 displays a persistent pattern of data and information falsification.

[45] ILCS field work took nearly two months so there is not one unambiguously correct IBC number to compare with.

4. CONCLUSION

In section 2 I measured L2 against the AAPOR (2005) and argued that there had been a number of violations of principles of professional responsibilities in dealing with respondents and in standards for minimal disclosure. In particular, it is likely that there were inadequacies in L2's informed consent processes and that respondents were endangered and their privacy was breached. The L2 authors have failed to disclose important information including the exact wordings of the questions that were asked, a definitive data-entry form, their full sample design and data matching anonymized interviewer IDs to households.

In Section 3, and also to some extent in Section 2, I presented evidence of data fabrication and falsification that includes:

1. Evidence suggesting that the figure of 600,000 violent deaths was extrapolated from two earlier surveys.

2. Shortcomings of disclosure just mentioned including the L2 questionnaire, data-entry form and sample design, and data that matches interviews with anonymized interviewer IDs.

3. Improbable response rates and success rates in visiting selected clusters despite highly insecure conditions.

4. Presence of many known risk factors for fabrication listed in AAPOR/ASA (2003).

5. A claimed field-work schedule that appears to be impossible, at least without committing ethical transgressions in the field.

6. Large discrepancies with other data sources on the scale, location and timing of violent deaths in Iraq in ways that are consistent with fabrication and the use of a trend figure (section 3.8) that hides these timing discrepancies.

7. Evidence of fabrication in a particular Baghdad cluster (cluster 33) combined with the implausible claim of zero security-related failures to visit Baghdad clusters during a period when Baghdad was very insecure and further evidence of fabrication in a cluster in Nineveh (cluster 34).

8. Unlikely patterns in the confirmations of violent deaths through the viewing of death certificates and in the patterns on when deaths certificates were requested and when they were not requested.

9. Manipulation of other evidence on mortality in Iraq and material that is not relevant to mortality in Iraq or unsuitable for citation in a scientific publication.

A few of these anomalies could occur by chance but it is extremely unlikely that all of them could have occurred randomly and simultaneously. In light of these findings, Burnham et al. (2006a) cannot be considered a reliable contribution to knowledge about mortality during the Iraq war.

I conclude that there should be a formal investigation of the second *Lancet* survey of mortality in Iraq. To aid such an investigation L2 authors should first meet the minimal disclosure standards established by AAPOR and, in addition, should provide access to their raw data, including the filled-out data-entry forms (anonymized if necessary) and sampling details.

Bibliography

AAPOR (2005) AAPOR code of professional ethics and practice. http://www.aapor.org/aaporcodeofethics.

AAPOR and ASA (2003) Interviewer fabrication in survey research. http://www.amstat.org/sections/SRMS/fabrication.pdf..

ABC (2007a) Ebbing hope in a landscape of loss marks a national survey of Iraq. http://abcnews.go.com/images/US/1033aIraqpoll.pdf.

ABC (2007b) Iraq poll: note on methodology. http://abcnews.go.com/US/story?id=3571535&page=1.

Ball, P., Kobrak, P. and Spirer, H.F. (1999) *State violence in Guatemala, 1960-1996: a quantitative reflection.* American Association for the Advancement of Science and Centro International Para Investigaciones en Derechos Humanos.

BBC (2006) Baghdad market blast kills scores. July 1, 2006, http://news.bbc.co.uk/1/hi/world/middle_east/5136028.stm.

Benini, A. and Moulton, L. (2004) Civilian victims in an asymmetrical conflict: operation enduring freedom, Afghanistan. *Journal of Peace Research* **41**(4), 403-422.

Biever, Celeste (2007) Winning the war for Iraq's dead. *New Scientist*, April 25, 2007.

Bohannon, J. (2006) Iraqi death estimates called too high: methods faulted. *Science* **314**(5798), 396 – 397.

Bohannon, J. (2008) Calculating Iraq's death toll: WHO study backs lower estimate. *Science* **319**(5861), 273.

Bloomberg School of Public Health (2007) Release of data from the 2006 Iraq mortality study. http://www.jhsph.edu/refugee/publications_tools/iraq/index.html.

Burkle, F.M., Tapp, C., Wilson, K., Takaro, T., Guyatt, G.H., Amad, H. and Mills, E.J., (2008) Iraq war mortality estimates: a systematic review. *Conflict and Health* **2**(1), March 7, 2008.

Burnham, G. (2007) Counting the dead in Iraq. Videotaped lecture given at MIT, http://mitworld.mit.edu/video/453/.

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

Burnham, G., Doocy, S., Dzeng, E., Lafta, R. and Roberts, L. (2006b) The human cost of the war in Iraq. MIT, http://web.mit.edu/cis/human-cost-war-101106.pdf.

Burnham, G., Lafta, R., Doocy, S. and Roberts, L. (2007) Authors' Reply. *The Lancet*, **369**(9556), 103-04.

Burnham, G. and Roberts, L (2006a) A debate over Iraqi death estimates. *Science* **314**(5803), 1241.

Burnham, G. and Roberts, L (2006b) Counting corpses: the *Lancet* number crunchers respond to *Slate's* Fred Kaplan. November 20, 2006, http://www.slate.com/id/2154203/?nav=navoa.

Burnham, G. and Roberts, L. (2008) The authors respond. *National Journal*, January 19, 2008. http://personal.rhul.ac.uk/uhte/014/Letter%20to%20the%20National%20Journal.pdf.

Checchi, F. and Roberts, L. (2005) Interpreting and using mortality data in humanitarian emergencies: a primer for non-epidemiologists. HPN Network Paper, no. 52.

Cole, J. (2007) Informed Comment. January 31, 2007, http://www.juancole.com/2007/01/bush-comment-on-najaf-farcical.html.

Department of Defense (DoD) (2006) Measuring stability and security in Iraq. August, 2006. http://www.defenselink.mil/pubs/pdfs/Security-Stabilty-ReportAug29r1.pdf.

Dardagan, H., Sloboda, J. and Dougherty, J. (2006a) Speculation is no substitute: a defence of Iraq Body Count. http://www.iraqbodycount.org/analysis/reference/pdf/a_defence_of_ibc.pdf.

Dardagan, H., Sloboda, J. and Dougherty, J., (2006b) Reality checks: some responses to the latest *Lancet* estimates. http://www.iraqbodycount.org/analysis/beyond/reality-checks/1

Deltoidblog (2006 and 2008) Les Roberts responds to Steven E. Moore, http://scienceblogs.com/deltoid/2006/10/les_roberts_responds_to_steven.php.

Dougherty, J. (2007) Mortality in Iraq. *The Lancet* **369**(9556), 102-103.

EPIC (2004) An interview with EPIC advisor Richard Garfield, http://www.epic-usa.org/An_Interview_with_EPIC_A.html.

Fafo (undated) Content of IMIRA (ILCS). http://www.fafo.no/ais/middeast/iraq/imira/content.htm.

Giles, J. (2007) Death toll in Iraq: survey team takes on its critics, *Nature* **446**(7131**)**, 6-7.

Goma Epidemiology Group (1995) Public health impact of Rwandan refugee crises: what happened in Goma Zaire in July, 1994? *The Lancet* 345(February 11), 339-44.

Gourley, S., Johnson, N., Onnela, J., Reinert, G and Spagat, M. (2007) The two *Lancet* surveys of Iraq do not validate each other. http://www.rhul.ac.uk/Economics/Research/conflict-analysis/iraq-mortality/L1_versus_L2.html.

Guha-Sapir, D., Degomme, O. and Pedersen, J. (2007) Mortality in Iraq. *The Lancet*, **369**(9556), 102.

Herold, M (2004) Daily casualty count of Afghan civilians killed by US bombing. http://pubpages.unh.edu/~mwherold/.

Hicks, M. (2006) Mortality after the 2003 invasion of Iraq: were valid and ethical field methods used in this survey? HiCN Research Design Note 3.

Iraq Body Count (continuously updated) http://www.iraqbodycount.org/.

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/ifhs.htm

Iraq Living Conditions Survey 2004 (ILCS) (2005a) Overview. http://reliefweb.int/rw/rwb.nsf/db900sid/KHII-6CC44A?OpenDocument

Iraq Living Conditions Survey 2004 (ILCS) (2005b) Volume I: tabulation report. http://reliefweb.int/rw/RWFiles2005.nsf/FilesByRWDocUNIDFileName/KHII-6CC44A-undp-irq-31dec1.pdf/$File/undp-irq-31dec1.pdf.

Janabi, A., (2006) Iraqi group: civilian toll over 37,000. July 31, 2004, http://english.aljazeera.net/archive/2004/07/200849155555897934.html.

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.

Kaiser, J. (2007) Iraq mortality study authors release data, but only to some. *Science* **316**(5823), 355.

Kane, D. (2007) The *Lancet* surveys of mortality in Iraq. http://cran.at.r-project.org/web/packages/lancet.iraqmortality/index.html.

Kuzman M., Tomic B. Stevanovic, R. et al. (1993) Fatalities in the war in Croatia, 1991 and 1992: underlying and external causes of death. *JAMA* 270(5), 626-28.

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

London School of Hygiene and Tropical Medicine (LSHTM) (undated) The use of epidemiological tools in conflict-affected populations: open-access educational resources for policy-makers. http://www.lshtm.ac.uk/hpu/conflict/epidemiology/index.htm.

Malfait, P., Moren, A., Dillon, J.C. et al. (1993) An outbreak of pellagra related to changes in dietary niacin among Mozambican refugees in Malawi. *International Journal of Epidemiology* **22**(3), 504-11.

Munro, N. and Canon, C. (2008) Data bomb. *National Journal*, January 4, 2008. http://news.nationaljournal.com/articles/databomb/index.htm.

Munro, N. (2008) Unscientific methods? *National Journal*, January 4, http://news.nationaljournal.com/articles/databomb/sidebar.htm#

NGO Coordination Committee of Iraq (2006) Iraq Emergency Situation. http://www.ncciraq.org/IMG/pdf/NCCI_-_Iraq_Emergency_Situation_-_Final_Report_-_2nd_May_2006.pdf.

O'Hanlon, M.E. and Kammons, A. (2006) Iraq index: tracking variables of reconstruction and security in post-Saddam Iraq. Washington DC: The Brookings Institution. http://www.brookings.edu/saban/iraq-index.aspx/.

ORB (2008) Update on Iraq Casualty Data. http://www.opinion.co.uk/Newsroom_details.aspx?NewsId=88.

Roberts, L. and Despines, M. (1999) Mortality in the Democratic Republic of the Congo. *The Lancet* **353**(9171), 2249-50.

Roberts, L, Hale, C., Belyakdoumi, F. et al. (2001) Mortality in eastern Democratic Republic of Congo. New York: International Rescue Committee, http://www.grandslacs.net/doc/3741.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. *The Lancet*, 364(9448), 1857-1864.

Roug, L. and Smith, D. (2006) War's Iraqi death toll tops 50,000. *Los Angeles Times*, June 25, 2006, http://www.commondreams.org/headlines06/0625-03.htm

Sloboda, J., Dardagan, H. and Bagnall, P. (2007) How can the utility of press reports be assessed? http://www.iraqbodycount.org/analysis/qa/assessment/.

SMART (2006) Measuring mortality, nutritional status, and food security in crisis situations: SMART METHODOLOGY. http://www.smartindicators.org/SMART_Methodology_08-07-2006.pdf.

Spagat, M. (2007) The discussion of possible sampling bias in the second *Lancet* study of mortality in Iraq. http://personal.rhul.ac.uk/uhte/014/Households%20in%20Conflict%202007.pdf.

Spagat, M. (2008) Counting the dead in Iraq. Presentation given at the JSM meetings in Denver, CO, August 6, 2008, http://personal.rhul.ac.uk/uhte/014/Denver.pdf.

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

Steele, J. and Goldenberg, S. (2008) What is the real death toll in Iraq ? *The Guardian* March 19, 2008.

The National Interest (radio program) (2006) Counting the Dead in Iraq, ABC Radio International, http://www.abc.net.au/rn/nationalinterest/stories/2006/1778810.htm.

UN Assistance Mission for Iraq (UNAMI) (2007) Human Rights Report: 1 November – 31 December 2006. http://www.uniraq.org/FileLib/misc/HR%20Report%20Nov%20Dec%202006%20EN.pdf.

United Press International (UPI) (2005) Iraqi civilian casualties, July 12, http://iraqmortality.org/iraqi-civilian-casualties.

United States State Department (2005) A trio of disinformers: *Islam Memo*, Muhammad Abu Nasr, and *Jihad Unspun*.
http://usinfo.state.gov/media/Archive/2005/Apr/08-205989.html

Wanniski, J. (2003) Civilian war deaths in Iraq.
http://personal.rhul.ac.uk/uhte/014/Wanniski%2037,000%20Dead.html