American Association for Public Opinion Research

THE POLLS—REVIEW

METHODOLOGICAL REVIEW OF "MORTALITY AFTER THE 2003 INVASION OF IRAQ: A CROSS-SECTIONAL CLUSTER SAMPLE SURVEY"

DAVID A. MARKER

Abstract   Burnham et al. (2006, *Lancet* 368:1421–28) described a household survey of Iraq that attempted to estimate the number of excess casualties since the invasion of that country in 2003. This review examines many of the key factors that could affect the accuracy of this estimate.

Introduction

Over the last two years, a wide variety of estimates have been circulated for the number of Iraqis who have died, directly or indirectly as a result of the invasion by the United States and other countries, and from the subsequent instability that continues to this day. Most of these estimates were the result of counts of dead who met a set of criteria developed by the counting organization. Burnham et al. (2006) provide an alternative estimate based on a national household survey of the country. Their estimate of 654,965 (95 percent confidence interval 392,979–942,636) is much larger than earlier counts for the postinvasion period of March 2003 through July 2006. In fact, their estimate corresponds to an average of 16,794 excess deaths per month, or 560 per day, throughout this period. This review attempts to clarify the methodological differences and identify the key assumptions underlying the estimates. We do not attempt to produce an alternative estimate; rather the goal is to identify the strengths and weaknesses of their approach so that potential users are able to judge its accuracy.[1]

1. The Burnham et al. study has been the object of much scrutiny and criticism. See, for example, Bohannon (2006, 2008); Giles (2007); Biever (2007); Kaiser (2007); Munro and Cannon (2008). This paper examines some of the issues that have been raised, as well as some others drawn from the perspective of survey quality.

DAVID A. MARKER is with Westat, 1650 Research Boulevard, Rockville, MD 20850, USA. The author wants to thank Graham Kalton and Ralph DiGaetano of Westat, along with the comments of the editor and eight referees for their suggestions. Address correspondence to David A. Marker; e-mail: DavidMarker@Westat.com

We begin with a brief overview of the Burnham et al. (2006) methodology. More details are provided in their article.

Groves et al. (2004) provide an excellent overview of survey quality in Chapter 2 of their book. The focus here is on what they refer to as representation, "What populations are described by the survey—who is the survey about?" We focus on four factors that can affect survey quality: coverage errors, correct probabilities of selection, effect of migration on estimation, and training and control of interviewers.

Finally, we use this review to identify how future attempts to estimate mortality during a violent conflict might be improved.

Overview of Burnham et al. Methodology

Burnham et al. conducted a national retrospective longitudinal study of Iraqi deaths from January 2002 through July 2006 (data were only collected at one time, but historical information was collected concerning multiple time periods). Households were asked about deaths during the 13 months preceding the invasion to serve as a baseline from which postinvasion excess deaths could be extrapolated (the number of deaths observed postinvasion compared to an assumed steady-state preinvasion rate). Exact date of death was requested for all household members who had died between January 1, 2002 and June 30, 2006. These deaths were allocated to the preinvasion time period (January 2002 to February 2003) or three similar-length postinvasion periods (March 2003 to April 2004, May 2004 to May 2005, and June 2005 to June 2006). Death certificates were requested at 87 percent of homes reporting at least one death, and were presented for 92 percent of those deaths.

Long recall periods such as these can introduce problems of assigning historical events to the correct time period (Neter and Waksberg 1964). Burnham et al. tried to minimize this problem by asking for the date of death and then coding these dates into the three time periods. The use of an important reference date, the invasion, is likely to minimize any deaths being confused between the before and after invasion time periods (Murphy and Cowan 1984). (This is similar to asking Americans about events occurring before or after the September 11, 2001 attack.) But it is quite possible that some of the deaths are assigned to the wrong one of the three postinvasion periods. This would only affect trend estimates, not estimates of overall number of deaths since the invasion. Trying to remember in 2006 whether a death occurred in late 2001 or early 2002 might be problematic, introducing some potential error into the baseline death rates. The authors examined this by comparing the death rates for the preinvasion period and the first two postinvasion periods from the current study and their earlier survey in Iraq during 2004 and found very similar rates for all the three periods. So, if there were some recall bias problems they were similar in two independent replications of the survey methodology. One additional important

recall error minimization technique was the use of death certificates, which presumably had dates of death.

Collecting this history of deaths can also be problematic because the household composition can change over the time period. Marriage, migration, and other changes can cause households to split, combine, or mutate. Suppose a new housing unit was constructed, and members of two or more former households merged or a previous household split. Then, this could result in undercounting (if the person providing the mortality data was unaware of someone who had died) or double counting (if both households reported someone who had died and been a member of the previous household). It is not clear how the authors dealt with changing households during this recall period (beyond the discussion below, in the section on coverage, of a three-month requirement).

A two-stage stratified sample was selected. Initially, 50 clusters were allocated to 18 strata representing the 18 governorates in proportion to the 2004 mid-year population estimates produced by the UNDP/Iraqi Ministry of Planning. Each governorate received at least 1 cluster, with Baghdad receiving 12. The first stage of sampling consisted of a probability proportional to the size selection of administrative units within governorates, again based on the estimated population.

The second stage was a selection of a cluster of 40 housing units. This was accomplished by a

> 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. On the residential street, houses were numbered and a start household was randomly selected. From this start household, the team proceeded to the adjacent residence until 40 households were surveyed. For this study, a household was defined as a unit that ate together, and had a separate entrance from the street or a separate apartment entrance. (Burnham et al., p. 1422)

A total of 1,849 households completed the survey.

The field team in all clusters consisted of four interviewers (two males and two females) who were medical doctors with previous survey and community medicine experience. They where fluent in English and Arabic and at least some were fluent in other languages spoken in Iraq.

The study received ethical approval from the Committee on Human Research of the Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, and the School of Medicine, Al Mustansiriya University, Baghdad, Iraq.

On two occasions, miscommunication about exact locations led to clusters not being visited and on the third, insecurity caused the field team to choose the next nearest population area (as allowed in the data collection protocol), which turned out to be in another governorate. As a result, only 47 clusters were used in the survey, and no data were collected from the PSUs (clusters) from the strata representing Muthanna and Dahuk provinces.

Table I. Estimated Iraqi Deaths (Civilian or Excess) since March 2003 Invasion

Source	Estimate	Reference period
Iraqi body count (British NGO)	43,491–48,283	As of September 26, 2006
Iraqi Ministry of the Interior	80,000	As of September 26, 2006
Iraqi Health Minister Ali Hussein al-Shamari	150,000	As of October, 2006
Iraqi NGO Iraqiyun	128,000	As of July, 2005 (none provided for most recent year)
Burnham et al. (2006)	392,979–942,636	As of June 30, 2006
Iraq Family Health Survey	104,000–223,000	As of June 30, 2006

SOURCE.—Burnham et al. 2006 except the Iraqi Ministry of the Interior estimate is from www.brookings.edu/iraqindex, the Iraqi Health Minister is from the *Washington Post* (2007), and Iraq Family Health Survey is from IFHS Study Group (2008).

Survey Quality

There are many factors that can affect survey quality. Biemer and Lyberg (2003); Groves (1989); and Groves et al. (2004) all provide excellent overviews of this topic. Four factors seem particularly relevant to the Burnham et al. methodology: coverage errors, correct probabilities of selection, migration, and training and control of interviewers.

Coverage errors result when the sampling frame does not completely cover all units in the target population (Groves et al. 2004). No matter what methods (e.g., sampling, secondary analysis of field reports) are used to collect data on the sampling frame, those units not included in the frame never enter into the estimates unless one extrapolates to external control totals.

This is particularly a concern for many of the competing estimates that have been discussed in the press. Table 1 shows a number of the estimates that have been released. These estimates are for the approximate period March 2003 through the summer or fall of 2006, although the exact periods covered are shown in the right-hand column.

To compare these numbers, it is necessary to examine what is covered by the estimates.

The Iraqi Body Count (43,491–48,283) is restricted to media-reported civilian deaths only. This includes civilian deaths caused by coalition military action and by military or paramilitary responses. It also includes excess civilian deaths caused by criminal action resulting from the breakdown in law and order (Iraq-BodyCount.org/background). There are many parts of the country where the media are not regularly present. The media have become more restricted as the

country has gotten more violent. Thus, it is reasonable to assume that many deaths are not covered by the methodology used by the Iraqi Body Count.

The Iraqi Ministry of the Interior (80,000) collects deaths primarily from police stations, police units, and emergency patrols. It does "not include the wounded who die later from their injuries, those kidnapped and later killed, armed men who die in clashes with the U.S. or Iraqi forces, unidentified bodies, and other categories of death" (Washington Post 2007). There have been many stories about the lack of a police presence in Anbar and other governorates, so deaths in these areas also are not likely to have been included in the Ministry of the Interior estimate.

The Iraqi Health Ministry (150,000) relies on death certificates from morgues and government hospitals. It includes Iraqis killed in bombings, terrorist acts, militia attacks, and other acts of violence. The Health Ministry is controlled by the party of Moqtada al-Sadr, whose Mahdi Army is behind many of the sectarian killings. Asher (2007) pointed out that given the strong desire of religious Muslims in Iraq to provide a proper burial, including internment within 24 hours, death certificates are often obtained from local doctors without going to morgues or government hospitals.

Each of the above methodologies uses a counting, rather than estimation, approach. Methods based on counting will provide underestimates. This is because when one counts, all deaths fall into one of three categories: confirmed deaths, reported but not confirmed, and not reported. Counting systems include only the first of these. Examples of the second category are deaths reported by unreliable press or local observers, whereas deaths occurring in remote areas may never get reported by anyone. Such counting methods are very useful in providing a floor for other estimates, but cannot be expected to produce unbiased estimates.

In contrast, coverage errors for the Burnham et al. estimate are smaller. The two provinces where no data were collected, Muthanna and Dahuk, were assumed to have no excess deaths. This is clearly a source of underestimation, but as will be discussed below, the magnitude of this error is likely to be only a few thousand. Their other source of undercoverage is that housing units with no one left living at the time of the survey were not included, and those deaths could not be included.

The recent Iraq Family Health Survey (IFHS Study Group 2008) is also a national survey and thus has similar coverage properties to Burnham et al. Due to violent conditions, they were not able to visit 10.6 percent of their sampled clusters. IFHS tried to adjust for this lack of coverage by using relative death rates reported by the Iraqi Body Count. They recognized that this "assumes that completeness of reporting for the IBC is similar for Baghdad and other high-mortality provinces" (p. 486). This is highly unlikely given the lack of credible media reports outside Baghdad. That the IBC does not fulfill this assumption is re-inforced by Table E-1 of the IFHS appendix, which shows great variability in IBC coverage, with its death rates 3.3 times as high as IFHS in Baghdad, 2.3

times higher in Anbar, 65 percent higher in other high-mortality governorates, and consistent with IFHS in low-mortality governorates (the latter two having few nonvisited IFHS clusters). It is not clear how much this coverage adjustment affects the IFHS estimate of 104,000–223,000 violent deaths from the invasion through June of 2006 (note that their estimate also does not include nonviolent deaths since the invasion).

Related to the coverage error is that Burnham et al. may not have recorded all deaths for a sampled household. "Deaths were recorded only if the decedent had lived in the household continuously for 3 months before the event." This was done to avoid double-counting deaths, but if someone moved households within three months of their death, they would have no chance of inclusion. It is also quite possible that respondents underreported deaths in the household, for example if they were worried about the retribution for participation in certain activities during which someone died. There are no estimates available for the magnitude of these last three sources of error.

To produce design-unbiased estimates it is necessary for all sampled units to be selected with known, nonzero probabilities of selection. Burnham et al. employed a stratified two-stage selection process. After deciding upon a sample allocation to the strata representing governorates, they selected clusters (administrative units) within strata and finally selected households within clusters. While the probability of selection at the first stage is easy to compute, the probability of selection at the latter stage is unknown.

To have known probabilities of selection at the last stage, one typically takes either a fixed n out of the N_i units listed in the cluster, or takes a fixed rate r from among the N_i. The former produces a constant workload for interviewers, but requires weighting to produce unbiased estimates, while the latter produces unbiased estimates without weighting but produces uneven workloads that can complicate the fieldwork. The authors did not follow such methodologies. The process used by the authors at this stage was not to estimate N_i, to take a fixed 40 occupied housing units, and not to weight the resulting data. This does not allow for design-unbiased estimation.

The authors used log-linear regression models of mortality rates and relative risks of mortality rather than design-based estimation. However, they did not adjust for the unequal weighting in this modeling, introducing potential biases in the estimation process. The direction of the bias cannot be determined since the appropriate weights are unknown. However, it is possible to describe some of this potential bias.

The population of Iraq was stratified based on the 18 governorates, and each of the 18 strata was allocated a number of clusters in proportion to its population, reflecting the mid-year 2004 population estimates for the governorates provided by the UNDP/Iraqi Ministry of Planning with a total population estimate of 27,139,584. Note that the actual number of clusters allocated to a stratum is a whole number, while that stratum's population represents a fractional part of
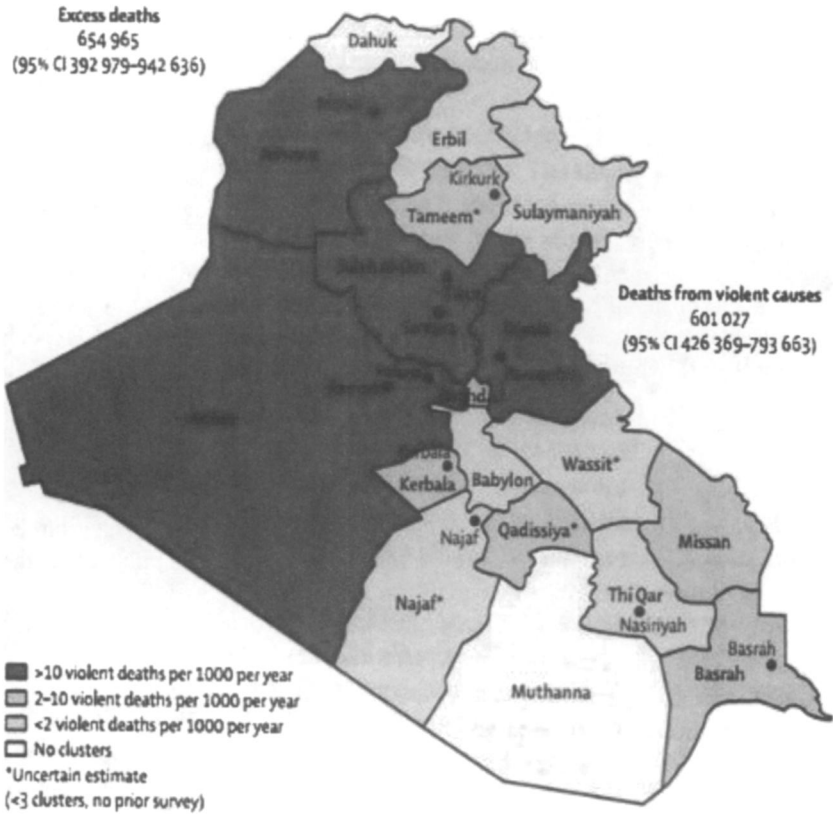
Figure 1. Map of Iraqi Excess Death Rates by Province from Burnham et al.

the country. So, for example, while Ninewa governorate merits 4.71 of the 50 clusters based on the relative sizes of the governorates, it would be assigned either 4 or 5 for sampling purposes. In the sample chosen by Burnham et al., it received 5.

Thus, the allocation process to explicit strata results in some departure from a strict proportionate distribution of clusters to the various strata. If the estimation process takes stratum-specific estimates and then combines them according to the stratum population, the resulting estimate is unbiased. However, if overall estimates are produced without adjusting for the specific allocation that was used, and if the variable of interest (e.g., death rate) varies across strata, it will be biased. The direction of the bias depends on the specific allocation. To examine this for the Burnham et al. actual sample, we grouped governorates by whether the observed death rate turned out to be high, moderate, or low. Figure 1 (figure 3 from Burnham et al. 1996) shows the death rates by governorate. For this analysis, we grouped Baghdad with the four high death rate governorates

and the two unmeasured provinces with the low death rate governorates. While the five high death rate governorates merit 23.86 of the 50 clusters in a proportionate allocation of clusters, they randomly received 25 of the 49 measured clusters (including the two "measured" as 0 deaths). Thus, an unweighted sample, as was the case for the analyses presented in this paper, overrepresents high death rate governorates. (This overrepresentation is still true if Baghdad is not included with the other four governorates that the authors reported as having the highest death rates.) The sample also overrepresented low death rate ones, while moderate death rate governorates received only 7 compared to a proportionate allocation that would have produced an expected 9.42. Note that because the difference in death rates between high and moderate is much greater than the difference between low and moderate, the overall impact is a disproportionately high contribution from the "high death rate" governorates. Without more information on the estimation process than is provided by the authors, it is impossible to determine how much this might have affected the estimates.

The decision to exclude sampled administrative units (clusters) that turned out to be in another governorate indicated a lack of understanding about the implementation of a sample design for a survey. The misclassification of units to one stratum instead of another creates no issues of bias for design-based estimation when the probabilities of selection are known. The data from misclassified sampled units can be appropriately reflected in the survey estimates through the use of sample weights. The penalty that is paid for the misclassification is an increase in variance, but that would be captured through the computing of standard errors reflecting the complex weighting design (using estimation methodologies such as the authors indicated they did in reflecting the concentration of the sample within 47 clusters).

It is important to consider how departures from an equal probability design arise with the authors' sample design. Within governorate i, n_i administrative units (the clusters) are selected, and then 40 households are selected per administrative unit. If we let T_i be the 2004 total population of the governorate, A_{ij} the 2004 population of the jth administrative unit in governorate i, and B_{ij} the 2006 number of households in the jth administrative unit at the time of data collection in 2006, then for each administrative unit j, we can compute the probability of selection of interviewed households:

$$\text{Prob(hhld)} = \text{Prob(Admin Unit)} * \text{Prob(hhld | Admin Unit)}$$
$$= (n_i * A_{ij} / T_i) * (40 / B_{ij}).$$

This only yields an equal probability sample within governorate i if $A_{ij} \propto B_{ij}$. There are several reasons why this may not be the case. First, the A_{ij} values are subject to error, both sampling and nonsampling with departures

between A_{ij} and B_{ij} arising as a result. Second, population distributions across administrative units can be expected to have changed since 2004, particularly in war zones. If the population in a given administrative unit shrinks faster in one unit than another, there will be fewer existing households from which to select 40 households than would be expected based on 2004 population figures, resulting in higher probabilities of selection in the areas experiencing a greater rate of reduction. Third, the measure of size A_{ij} was in terms of people but the values B_{ij} are in terms of households (occupied housing units). The number of people per household likely varies across all administrative units. It is quite likely that the average household size for Kurdish households is different than that for Shiites or Sunnis. Similarly, rural households may typically be larger (or smaller) than urban households. If a unit has larger average household sizes, it will have a smaller number of households and thus each one will have a greater chance of selection. Treating these unequal probabilities of selection as equal, as was done by Burnham et al., can introduce a bias to the estimates. It is impossible to determine the direction or magnitude of this source of bias for the particular sample that was selected; however, Burnham (private communication) did indicate that the average household size in low death rate governorates was about 9 percent smaller than that in moderate and high death rate governorates.

In addition, differential probabilities of selection are a source of variability in design-based estimates. This source of variation was ignored in the estimates of variability of study estimates presented by the authors. As a result, it can be expected that the confidence intervals provided by the authors are narrower than they should have been.

Burnham et al. recognized that migration could affect the quality of their estimate. "The population data used for cluster selection were at least 2 years old, and if populations subsequently migrated from areas of high mortality to those with low mortality, the sample might have overrepresented the high-mortality areas" (p. 1427). The United Nation's High Commissioner for Refugees (UNHCR) estimated that since the war began in March 2003, 1.6 million Iraqis have been displaced internally, and up to 1.8 million are living outside the country (Washington Post 2006). In November 2007, the UNHCR revised the estimate of internally displaced Iraqis since the invasion to 1.4 million, in addition to 1.0 million displaced internally before the invasion (UNHCR 2007).

If 1.8 million fewer Iraqis live in the country than assumed by Burnham et al., it will definitely result in an overestimate of mortality. If they depart more frequently from violent provinces, the overestimate will be even greater. If, for example, one assumes the average *external* migration rate (6.7 percent) for moderately violent provinces, a 50 percent higher rate (10 percent) in the most violent provinces, and a 2 percent rate in the least violent, then their estimate will be reduced by 10 percent.

One more point is worth making regarding external migration. Burnham et al. inflate the observed death rate to the population including those who had

emigrated at the time of the survey. This has the effect of assuming that house-holds that fled the country had similar levels of death as those that remained in the country. Excluding the 1.8 million, as is done in table 2, assumes that those households fled before anyone died, which is clearly an underestimate. The survey cannot provide direct evidence of death rates among the emigrant households, since by definition they were not in Iraq to be surveyed.

Figure 1 shows a map of the death rates found by Burnham et al. for each of the Iraqi governorates. As an example of the possible impact of internal migration, we look at what happens if an additional 1.4 million Iraqis moved from the four most violent governorates (Anbar, Diyala, Ninewa, and Salah al-Din) and Baghdad to the rest of Iraq. Table 2 shows the 2004 population estimates used by Burnham et al. in the first column, followed by the reduced population based on *external* migration (allocated as assumed above). The next column shows the effect on population of the assumed constant *internal* migration rate from five governorates to the others. The death rates in the fourth column are the midpoints of the ranges shown in figure 1, along with an average rate for the five most violent governorates that yields the approximate total excess deaths reported in their paper. Under these assumptions, the *combined external and internal* migration reduces their estimate by almost 20 percent. It is important to recognize that alternative allocations of internal migration would yield other estimates, either larger or smaller, but this one estimate is used simply to provide insight into how sensitive the reported estimates are to these assumptions. (For example, IFHS assumed between 350,000 and 600,000 external migrants (IFHS Table E-2).) Further, it is important to note that the UNHCR estimate was for migration since the war began in March 2003, while the population estimates used by Burnham et al. were from mid-year 2004. Depending on the accuracy of these two sets of figures, internal or external migration that occurred in the first 15 months after the invasion may have already been captured in the Burnham et al. estimate (the UNHCR (2007) estimated that only 15 percent of the internal displacement took place between 2003 and 2005). If this is the case, one would expect the effect of migration to be reduced compared to that shown in table 2. Similarly, if one includes an estimate for deaths among households that emigrated, it will again reduce the effect.

The data in table 2 make two assumptions not included in Burnham et al. First, we have assumed that Baghdad should be included in the most violent governorates, even though their finding was that it fell in the moderate range. Thus, table 2 assumes that Baghdad is one of the areas from which people are fleeing as part of internal migration. To strictly rely on the findings of Burnham et al., Baghdad should be one of the net receiving areas. This restricts the net 1.4 million internal migrants to come from four governorates, which contain only half the initial population used in table 2. To achieve the same observed overall number of deaths (650,000) thus requires a much higher excess death rate of 9 percent in these governorates. The internal migration has a more

Table 2. Sensitivity Analysis of External and Internal Migration

Province	2004 population	External migration	Internal migration	Estimated excess death rate (%)	Estimated excess deaths	Migration-adjusted deaths
Baghdad	6,554,126	5,902,084	5,193,455	4.73	310,010	245,650
Ninewa	2,554,270	2,300,157	2,023,990	4.73	120,817	95,735
Basrah	1,797,758	1,678,524	1,850,313	0.60	10,787	11,102
Sulamaniyah	1,715,585	1,682,958	1,855,201	0.10	1,716	1,855
Thi-Qar	1,493,781	1,465,372	1,615,347	0.10	1,494	1,615
Babylon	1,472,405	1,444,403	1,592,231	0.10	1,472	1,592
Erbil	1,418,455	1,391,479	1,533,890	0.10	1,418	1,534
Diyala	1,392,093	1,253,600	1,103,087	4.73	65,846	52,176
Anbar	1,328,776	1,196,582	1,052,915	4.73	62,851	49,803
Salah al-Din	1,119,369	1,008,008	886,982	4.73	52,946	41,954
Najaf	978,400	959,793	1,058,023	0.10	978	1,058
Wassit	971,280	952,808	1,050,324	0.10	971	1,050
Qadissiya	911,640	851,177	938,291	0.60	5,470	5,630
Tameem	854,470	797,798	879,449	0.60	5,127	5,277
Missan	787,072	734,870	810,081	0.60	4,722	4,860
Kerbala	762,872	712,275	785,174	0.60	4,577	4,711
Muthanna	554,994	544,439	600,160	0.00	0	2,101
Dahuk	472,238	463,257	510,669	0.00	0	511
Total	27,139,584	25,339,584	25,339,584		651,203	528,214

dramatic effect on excess deaths in this situation, making the *combined external and internal* migration reduction in deaths 30 percent, rather than 20 percent.

The second assumption is that internal migrants flow to all parts of the country that are more peaceful. It has been pointed out that only Kurds are likely to migrate to Kurdistan, and since they probably comprise a relatively small number of the internal migrants, table 2 should not show the four Kurdish governorates receiving many migrants. Redoing these computations with all internal migrants staying in Shiite or Sunni governorates only increases the estimated number of deaths by less than 1,000, so this does not have a major impact on the estimates.

The quality of survey data ultimately is dependent upon the interaction between interviewers and respondents. It is vital for unbiased estimation that the study staff identify the housing units eligible for sample selection (e.g., via lists) and then interviewers select respondents in a manner that provides every respondent a known, nonzero probability of selection. In a war zone, this activity can be extremely dangerous and difficult to control. Nevertheless, it is vital that such activities be controlled to the extent possible as they represent an important potential source of nonsampling error.

The America-based authors could not observe training or data collection in Iraq in 2006. (One of the authors did observe the data collection in 2004 for their earlier study.) They met with their Iraqi coauthor in advance, who then conducted the training in Iraq. Unfortunately data were not tracked by individual data collector, nor was data collection observed in relatively more peaceful areas, so there is no independent documentation of the quality of data collection.

In contrast, IFHS trained all central and local supervisors in Amman Jordon, after which the central supervisors trained interviewers in Iraq. Data entry staff received one-week training and survey instruments were pilot tested in all governorates. Field teams were closely supervised and each week, completed clusters were checked and incorrect or incomplete forms were returned.

A few years ago, 35 leading survey researchers issued a consensus statement on how to minimize interviewer falsification of data (AAPOR 2003). This statement has been endorsed by the American Association for Public Opinion Research and the Survey Research Methods Section of the American Statistical Association. They listed eight factors that could affect falsification rates. Inadequate supervision, poor quality control and off-site isolation of interviewers were three of those factors that are present in this study. The remaining five factors (training on falsification, interviewer motivation, inadequate compensation, piece-rate compensation, and excessive workload) are harder to assess in this situation due to the limited information available on these topics. When collecting data on controversial topics, it is very important that steps be taken (and documented) to avoid falsification so that those who disagree with the findings cannot use this to try to discredit them.

Burnham et al. report that in an earlier survey in Iraq in 2004, the sampled clusters were selected using GPS locators. By the time of the 2006 survey, this was not considered practical, since walking around a neighborhood carrying a GPS locator might appear to some to resemble those placing roadside explosive devices. This modification of survey procedures (and others reported by the authors) to reduce potential harm to data collectors is appropriate and laudable.

The procedure implemented eliminated the identification of housing units eligible for sample selection completely. A main street was randomly selected within the chosen administrative unit, and then a residential side street was selected that intersected the main street. Finally, a random starting house was selected on that street and then 40 adjacent occupied housing units were interviewed. The selection of streets and starting house were based on aerial photographs. Unfortunately, streets have different numbers of housing units. Moreover, side streets can cross multiple main streets, resulting in side streets having unequal chances of selection. For these and other reasons, the process employed to determine a starting point for household interviewing resulted in unknown and unequal probabilities of selection that represent a potential source of bias and add to the variability of the estimates. It should be noted that there are straightforward procedures regularly used by survey research organizations to sample housing units within sampled clusters that avoid the problems and issues introduced by the procedures used by the authors. These could have been employed without jeopardizing the safety of the interviewers.

Burnham et al. used a design based on the World Health Organization's Expanded Programme on Immunization (EPI) cluster survey design (WHO 1991). The EPI was created to allow for inexpensive surveys of immunization rates in areas where lists of households do not exist. This approach has been strongly criticized by the survey statistical community for more than 10 years, because it does not provide known probabilities of selection and therefore cannot produce unbiased estimates. In many of the situations found in the developing world, it is likely that the EPI estimates of proportions are reasonable and cost-effective. But a variety of alternatives for how to select households within clusters have been proposed that do not dramatically increase costs while providing for unbiased estimation through the use of weights. These are to be preferred in almost all situations, especially when totals are being estimated. Examples include Milligan, Njie, and Bennett (2004); Turner, Magnani, and Shuaib (1996); Scott (1993); and Lepkowski (1993). Proposed improvements have included selecting compact clusters in list and use of aerial photographs. Lepkowski expressed concern that "the poor practice of the EPI simple cluster sampling method is now being used as a standard for inexpensive surveys on other health topics." This concern is demonstrated here by its use for estimating mortality.

Turner, Magnani, and Shuaib (1996) deplore the inability of the EPI method to produce (design) unbiased estimates, pointing out the "need to maintain sufficient scientific rigor in the conduct of sample surveys such that the resulting

data may be confidently used to make important policy and program deci-
sions." The latter point goes to the heart of the argument against using the EPI
methodology for such high-profile policy evaluations as deaths in Iraq.

This lack of a known probability of selection was compounded by the need to
allow interviewers some latitude in avoiding areas they considered unsafe. As a
result, the level of control of housing units left to interviewers could introduce
bias. This concern is demonstrated by the reported contact and response rates in
the paper. They reported achieving 1,849 completes with only 16 not at home
and 15 refusals. This represents a 98.4 percent response rate and a 99.1 percent
contact rate. Given the safety concerns in Iraq, it is plausible that there is almost
always someone at home, but for the same reasons it is quite surprising that
almost everyone was willing to participate in the survey.

In considering these rates, it is important to recognize how different response
and contact rates can be in different parts of the world. Surveys in the developed
world do not achieve rates close to these numbers. On the other hand, refer-
ees have indicated that they regularly observe similar rates in the developing
world. The danger in a war zone may be counterbalanced by the interest in
communicating with someone actually interested in what the household has to
say.

It is interesting to compare these reported response rates and contact rates
with those reported for an ABC News poll conducted in Iraq in early 2007
(http://ABCNews.go.com/US/story?id=2954886). They reported a 56 percent
response rate and a 90 percent contact rate. (The IFHS response rate was 85 per-
cent, including gaining participation from 96 percent of eligible households.)
Some of ABC's lower response rate is due to their selection of a random adult
respondent, where Burnham et al. appear to have allowed any adult to respond.
But it is also important to realize that ABC had a much higher level of control
over interviewing, with just over half of interviews "checked by supervisors
– 28 percent by direct observation, 14 percent by revisits, and 10 percent by
phone."

OUTSTANDING METHODOLOGICAL ISSUES

There are a few additional methodological issues that deserve comment. Some
in the press have criticized Burnham et al. for including a random sample
of only 50 clusters. This included an opinion article appearing in the Wall
Street Journal (Moore 2006). But using a small number of clusters doesn't
introduce bias; it simply increases the width of the sampling error confidence
intervals. In an active war zone, it is understandable that the authors were
cautious about how many sample locations to visit. (In contrast, the ABC
News poll spread a similar-sized sample of 2,212 completed interviews across
458 areas and IFHS spread their 9,345 interviews across 1,086 clusters of
10 households each.) It is vital that this clustering be reflected in the method

used to compute variances and confidence intervals. The bootstrap methods and regression modeling using generalized estimating equations mentioned by Burnham et al. appear reasonable, although more detail on this issue would have been helpful. It is important to recognize that the confidence interval presented by the authors (392,979–942,636) only reflects sampling error, the variability of the estimates. It does not include any of the possible biases that may have been introduced into the survey estimates that were identified earlier in this review.

Burnham et al. used 50 clusters of 40 households. The cluster size is quite large, which can result in very inefficient samples if people living near to each other have similar experiences. This is definitely a concern in the Iraqi situation where whole neighborhoods have been reported to suffer violent periods while other neighborhoods have remained peaceful. Burnham et al. improved on the EPI method by attempting to randomly select a starting point from throughout the cluster, compared to the traditional method of going to the center of the cluster and spinning a bottle to determine a starting location. But alternatives exist that both randomly select a starting point and identify the probabilities of selection for all units.

The questionnaire asked for the dates of death for anyone who lived in the household for at least three months before the time of their death. This procedure was meant to avoid double counting a death in two households, but it excluded deaths of all those who had switched households within three months of their death. It is not clear how often this happened. One could imagine a war-torn area where people felt threatened and moved in with relatives or neighbors, but were still killed. This decision to minimize double counting introduced a downward bias of unknown magnitude. On the other hand, it is also quite possible that respondents underreported deaths in the household, for example if they were worried about retribution for participation in certain activities during which someone died. No estimates of either of these potential biases are available.

Preinvasion death rate estimates for Iraq range from approximately 5 per thousand (U.S. Central Intelligence Agency) to 10 per thousand (U.S. State Department) (Zeger and Johnson 2007). Some have criticized Burnham et al. for their preinvasion estimate of 5.5 per thousand, arguing that by underestimating the preinvasion level, they are overestimating the excess death rate since the invasion. But Burnham et al. used a longitudinal survey, where the same households were asked about pre- and postinvasion deaths. So if by chance they sampled households that experienced lower deaths, this same sample was used in both time periods. There is always a danger that sampled households preinvasion were less likely to experience death in the household but their situations changed to make them more likely than the typical household postinvasion. But it is also possible that they switched in the other direction. By collecting the before and after data from the same households they attempted to minimize this problem from affecting their estimates.

Table 3. Deaths per 13-Month Period

| | | Adults (15–59) | | | |
	Children (0–14)	Men	Women	Elderly (60+)	All deaths
Preinvasion					
Nonviolent	14.0	19.0	6.0	40.0	80.0
Violent	0.0	1.0	0.0	1.0	2.0
Total	14.0	20.0	6.0	41.0	82.0
Postinvasion					
Nonviolent	13.3	12.3	13.0	42.0	82.3
Violent	8.7	78.3	5.0	3.7	100.0
Total	22.0	90.7	18.0	45.7	182.3

NOTE.—Age was reported for 610 of 629 deaths.

Table 3 examines the pre- and postinvasion deaths reported by sampled households. Numbers in this table (derived from Burnham et al.'s table 2) record the number of preinvasion deaths reported by the sampled households, while postinvasion numbers are the average of the three postinvasion periods. This makes both periods approximately 13 months in length. The table shows that the number of nonviolent deaths is very similar in both periods. It is interesting to note that the nonviolent death rate for men has gone down by a third, possibly due to many instead dying violent deaths. But this has been countered by the doubling of the nonviolent deaths among women. This is possibly due to the deteriorating health and safety conditions in Iraq. For example, the number of doctors and open hospitals has dropped dramatically, with many reports of women not being able to get to a hospital to deliver a baby.

By comparison, IFHS has a much lower preinvasion death rate of 3.17, below both U.S. government estimates. More surprisingly, while Burnham, IBC, and many other sources indicate that Iraq became much more violent in 2005 and 2006 than it was shortly after the invasion, IFHS found flat violent (and overall) death rates for each of the three postinvasion time periods.

Recommendations for Future Studies

Well-designed and monitored surveys are generally going to provide better estimates than counting systems in war zones. The mechanisms relied on to count all occurrences that break down in wartime, even if they worked well in earlier periods. Counting systems (e.g., Iraqi Body Count and Iraqi governmental agencies) provide useful lower bounds against which survey estimates can be compared. The problems Burnham et al. faced in attempting to conduct a high-quality survey in a war zone are not to be understated. Nonetheless, these are problems that survey researchers (both statisticians and methodologists)

regularly address. While established procedures would have to be adapted to deal with the realities of a war zone, this could be done while minimizing errors and using known probabilities of selection that allow for unbiased estimation.

- Geographic information systems and high-resolution graphics (e.g., Google-earth) can be combined with basic area-probability survey techniques to assure that known probabilities of selection are used.
- To the extent that financing will allow, attempts should be made to increase the number of sampled clusters and decrease the number of completes per cluster. This will increase the accuracy of survey estimates and allow for greater quality control checks on the reported data.
- Random selection of households can be based on those same graphics to assure appropriate sampling and allow for estimating changes since the time of sampling frame construction (e.g. migration).
- Weights can be incorporated into the estimation process that accounts for these probabilities and sample frame deficiencies.
- Interviewer reporting and tracking systems can be used that will increase user comfort with the field staff procedures, while still maintaining the required confidentiality and safety concerns. Interpenetrated interviewer assignments (Mahalanobis 1946) would allow for quality control of recorded data. Result codes should provide enough detail to record response rates according to AAPOR standards. Debriefing sessions with interviewers can be held after the first few data collection trips to identify successful and unsuccessful procedures. Re-interviews and other supervisor checks can be incorporated, at least in the relatively peaceful areas.
- Point estimation and variance estimation methodologies need to be carefully documented to make sure that they are appropriate for the complexities that arise in difficult situations.

Conclusions

Burnham et al. attempt to estimate the number of excess Iraqi war dead through-out the country using fairly standard survey methodology, for which they are to be commended. We have examined four specific methodological factors: coverage errors, correct probabilities of selection, migration, and training and control of interviewers. Coverage provided by the first stage of sample selection (the sample of administrative units) appears to have been complete (other than the exclusion of area considered as too violent to allow household interviews as well as the exclusion of the three sampled clusters for the reasons the authors indicated). The coverage at the second stage appears less than complete. In addition to the exclusion of some deaths mentioned earlier (short term house-hold members), there may have been systematic exclusion of some types of households or housing units. The extraordinarily high response rates reported suggest this as a possibility. It should be noted that for the two governorates

where data were not collected, they used an underestimate of no excess deaths, choosing to err on the side of understating mortality in these unknown cases. This level of coverage represents a major improvement over other reported estimates.

The procedures they used, unfortunately, do not allow the computation of their probabilities of selection, thus making it impossible to produce design-unbiased estimates. The authors treated their data as if it were obtained from an equal probability sample when they were not. Not only does this result in an underestimation of sampling error, but it also prevents them from adjusting for random over- or underrepresentation of violent/peaceful areas of the country.

Iraq has clearly experienced significant migration since the time of the invasion. The authors take account of migration from the March 2003 invasion through mid-2004, using the best population estimates at that time. However, as sectarian violence and death rates have increased from that time through mid-2006, the rate of external and internal migration has increased. It is likely that not accounting for this has produced an overestimate of the number of excess deaths.

Given the war zone in which data collection took place, the authors needed to develop specialized data collection procedures. This put extra reliance on the training and control of interviewers. Burnham et al. wisely modified procedures to improve the safety of their field staff. However, the study reports of almost perfect contact and response rates raise some question about how well procedures designed to obtain high-quality data were followed. Without more extensive data on the field operations, it is impossible to determine the potential impact of these actions; however, it is easy to imagine situations in which it could lead to either over- or underestimation.

We should not let these difficulties stand in the way of future attempts to use survey methods in difficult situations. Whether in a war zone or following a tsunami, collecting accurate information on the extent of the catastrophe and an accurate reporting of the types of assistance needed by the affected populace can only be accomplished through surveys. The authors of future reports will be attacked for their estimates, because no matter what they report, it will contradict the preconceived notions of someone. To quote sociologist Peter Rossi (1987), "No good applied social research goes unpunished." That is why it is so important that sound survey, sample design, and estimation methodologies be employed to permit the results of a study to be readily and credibly defended.

References

American Association for Public Opinion Research. 2003. Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of its Effects. Available at http://www.aapor.org/uploads/falsification.pdf.

Asher, Jana. 2007. "Discussion of 'Mortality after the 2003 Invasion of Iraq: A Cross-Sectional Cluster Sample Survey.'" Presented to the Washington Statistical Society and American Academy for the Advancement of Science, February 6, 2007.

Biemer, Paul, and Lars Lyberg. 2003. *Introduction to Survey Quality.* New York: Wiley.

Burnham, Gilbert, Riyadh Lafta, Shannon Doocy, and Les Roberts. 2006. "Mortality after the 2003 Invasion of Iraq: A Cross-Sectional Cluster Sample Survey." *Lancet* 368:1421–28.

Groves, Robert M. 1989. *Survey Errors and Survey Costs.* New York: Wiley.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology.* New York: Wiley.

Iraq Family Health Survey Study Group. 2008. "Violence-Related Mortality in Iraq from 2002 to 2006." *New England Journal of Medicine* 358:484–93.

Lepkowski, James M. 1993. "Discussion of Papers by Pember and Banda and by Bennett." *Proceedings of the 49th Session of the ISI, Bulletin of the International Statistical Institute* 4:97–98.

Mahalanobis, Prasanta C. 1946. "Recent Experiments in Statistical Sampling in the Indian Statistical Institute." *Journal of the Royal Statistical Society* 109:325–70.

Milligan, Paul, Alpha Njie, and Steve Bennett. 2004. "Comparison of Two Cluster Sampling Methods for Health Surveys in Developing Countries." *International Journal of Epidemiology* 33:469–76.

Moore, Steven E. "655,000 War Dead? A Bogus Study on Iraq Casualties." *Wall Street Journal,* October 18, 2006, http://opinionjournal.com/editorial/feature.html?id=110009108 (accessed February 28, 2008).

Munro, Neil. "Unscientific Methods?" *National Journal,* January 4, 2008, http://news.nationaljournal.com/articles/databomb/sidebar2.htm (accessed February 28, 2008).

Munro, Neil and Carl M. Cannon. "Data Bomb." *National Journal,* January 4, 2008, http://news.nationaljournal.com/articles/databomb/index.htm (accessed February 28, 2008).

Murphy, Linda, and Charles Cowan. 1984. "Effects of Bounding on Telescoping in the National Crime Survey." In *The National Crime Survey Working Papers (Vol II),* eds. R. Lehnen, and W. Skogan. Washington, DC: U.S. Department of Justice.

Neter, John and Joseph Waksberg. 1964. "A Study of Response Errors in Expenditures Data from Household Interviews." *Journal of the American Statistical Association* 59:18–55.

Rossi, Peter. 1987. "No Good Applied Social Research Goes Unpunished." *Society* 25:74–79.

Scott, Chris. 1993. "Discussion of Papers by Pember and Banda and by Bennett," *Proceedings of the 49th Session of the ISI, Bulletin of the International Statistical Institute* 4:95–97.

Turner, Anthony G., Robert J. Magnani and Muhammad Shuaib. 1996. "A Not Quite as Quick but much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design." *International Journal of Epidemiology* 25:198–203.

United Nations High Commission for Refugees. 2007. Iraq: UNHCR Cautious about Returns, at http://www.unhcr.org/cgi-bin/texis/vtx/news/opendoc.htm?tbl = NEWS&id = 4746da102 (accessed December 6, 2007).

Washington Post. 2007. "War's Toll on Iraqis put at 22,950 in '06." January 8, 2007, p. A1.

Washington Post. 2006. "1,000 Iraqis a Day Flee Violence, U.N. Group Finds." November 24, 2006, p. A13.

World Health Organization. Training for Mid-level managers: The EPI coverage survey. Geneva: WHO Expanded Programme on Immunization, 1991. WHO/EPI/MLM/91.10.

Zeger, Scott and Elizabeth Johnson. 2007. "Estimating Excess Deaths in Iraq since the US-British Led Invasion.'" *Significance* 54–59.