Chapter 11

The Guatemalan Commission for Historical Clarification: *Generating Analytic Reports*

Inter-Sample Analysis

Patrick Ball

Introduction

This paper reports on an analytical study requested by the Commission for Historical Clarification (CEH) and carried out by the American Association for the Advancement of Science.[1] The purpose of the study was to answer the question: How many people were killed in Guatemala during the period of the CEH mandate, 1960-1996? To answer this question, we used the information in three databases of human rights violations, resulting from three projects – one conducted by the CEH, one by the CIIDH, and one by REMHI. These databases reported data from interviews with direct witnesses and victims. As a consequence of having three sources, we must first ask a) how many unduplicated killings are documented by the three projects? and then attempt to answer the second question, b) how many killings were there in all during the internal armed conflict?

Our analysis deals with these two questions. We deal first with the information collected by the three projects in light of the objectives of this analysis. We then explain the scientific methods used to estimate rates and quantities that answer the specific empirical questions derived from these objectives. We then present and interpret the results of applying the scientific methods to the information from the three databases. We subsequently analyze four regions of Guatemala in which genocide may have occurred during the period 1981-1983. Finally, using other statistical methods we show that the three projects lead to similar implications about the patterns of violence in Guatemala.

Note that in some tables, there are numbers that are not counts, but result from arithmetic operations subject to rounding error. Thus, totals shown in the table will not exactly add up to the totals of the related columns or rows. In some graphs we have retained the Spanish labels, as it is our intent to present tables and graphs as they appeared in the CEH report.[2]

The Information

The three databases were created using information gathered from interviews with witnesses and victims. Each contained a list of named victims who were killed, as well as numbers of people who were killed but who could not be identified by name. The three projects did not define "political killing" in the same ways. Therefore the measure we use in this study is *deaths*, and not the more juridically precise term "extrajudicial execution" that is used elsewhere in the CEH report. The three projects have unique definitions of murder, and to join them, it was necessary to use the broadest possible definition of the killing as a human rights violation. Thus, the totals of deaths in the AAAS study should be compared with the totals of deaths in the CEH report, and not with the totals of extrajudicial execution.

Table 1 shows the number of documented killings (victims with and without names), by time period, region, and database. Many killings were not reported to any project, and therefore, the quantities presented in Table 1 are less than the total actual number of victims who were killed in political violence in the period 1960 to 1996. Table 1 shows only those victims who were reported to one or more documentation project.

[1] At the request of the CEH, this analysis was conducted by Dr. Patrick Ball, Deputy Director of the Science and Human Rights Program of the American Association for the Advancement of Science, with the assistance and advice of Dr. Herbert F. Spirer (Adjunct Professor, Columbia University and consultant to the AAAS), Dr. Frederick Scheuren (Senior Fellow of The Urban Institute, and Adjunct Professor at George Washington University), and William Seltzer (Senior Researcher at Fordham University).

[2] The meaning of the Spanish labels will be clear to any reader.

Table 1[3]: Number of documented killings (victims with and without names), by time period, region, and project

Region	CEH			CIIDH[4]			REMHI		
	1960-1977	1978-1996	1960-1996	1960-1977	1978-1996	1960-1996	1960-1977	1978-1996	1960-1996
Region 0 (others)	271	9,916	10,187	18	2,586	2,604	84	6,888	6,972
Region I (area Ixil)	14	4,609	4,623	0	4,028	4,028	9	5,423	5,432
Region II (Cahabón)	1	532	533	0	135	135	7	453	460
Region III (Rabinal)	0	1,379	1,379	0	1,297	1,297	0	1,354	1,354
Region IV (San Martín Jilo.)	0	1,347	1,347	1	20	21	0	68	68
Region V (Nte. De Huehue)	0	1,746	1,746	0	1	1	0	1,032	1,32
Region VI (área Zacualpa)[5]	0	1,951	1,951	0	238	238	1	1,674	1,675
Region VII (Guatemala)	91	245	336	1	15	16	10	111	121
Region VIII (área Panzós)	0	169	169	11	41	52	1	1,167	1,168
Region IX (Ixcán)	3	2,421	2,424	0	48	48	5	2,751	2,756
Region X (area Costa Sur)	25	190	215	2	91	93	23	139	162
Total	405	24,505	24,910	33	8,500	8,533	140	21,060	21,200

The three projects did not equally cover all of the regions. All conducted investigations in the Ixil area (Region I) during the period 1978-1996, but only the CEH collected adequate information in San Martín Jilotepeque (Region IV)[6]. It is also clear that none of the projects adequately covered

[3] Table 1 excludes the victims for whom the year or place of death is not known.

[4]. Although the CIIDH also collected information from journalistic and documentary sources, this analysis only includes the information from direct sources supported by the witness' signature.

[5] The definition of Region VI (the Zacualpa area) includes the municipios of Chiche and Joyabaj, and therefore does not correspond exactly to the definition of the region in the section of the CEH report that examines genocide. That section includes as Region VI only the municipio of Zacualpa.

[6] The regions were defined in order to isolate areas in which there were big differences in the coverage rates among projects.

the period 1960-1977, including the massacres of the 1968-1973. Any estimate must take these limitations into account.

If no victims were reported to more than one project, the total of documented victims would be the sum of the three totals, that is 24,910+8,533+21,200 = 54,643. However, many of the same victims were reported to two or three projects. Thus, we cannot assume that the total number of victims is equal to this simple sum.

The projects were managed independently, and each victim could have been reported to more than one project. For example, assume that a victim Juan Pérez was murdered. His wife may have reported the killing to the CIIDH in 1994; his son may have given testimony to REMHI in 1996; and Peréz's neighbor might have related the story to the CEH in 1997. If we simply sum the three databases, Peréz's killing will be counted three times.

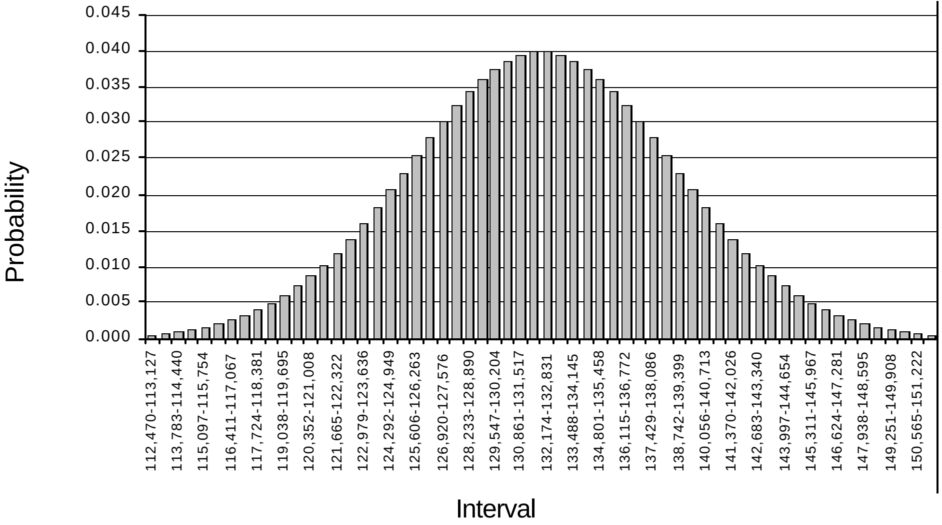
Duplicated reporting of deaths in more than one database is called "overlap." To estimate the total number of victims reported by the three databases, the overlap between databases must be estimated to reduce the contribution of each database by its overlap rate.

Two possible conditions demonstrate the limits of the overlap problem. If none of the victims in any database appear in any other database, then the total number of victims of killing and disappearance is equal to the sum of the number of victims in the three databases (54,643). This is the upper limit to the number of such victims. The lower limit can be found in the extreme case that the largest of the three databases (here, the CEH database) contains all the cases reported in the other two (REMHI and CIIDH). In this case, the total number of killings is simply the number of killings reported in the largest database (405+24,505=24,910). The sum of the total number of unique victims in the three databases must fall within these two limits, that is, between 24,910 and 54,643. The purpose of our analyses estimating the total number of documented killings is to narrow this range.

Many killings were not reported to any of the three projects. In the following section, we carry out a scientific analysis to estimate the total number of killings, [7] including those not reported to CEH, the CIIDH, nor to REMHI. The estimate from this analysis is that between 119,300 and 145,000 killings were committed, with the most likely figure being around 132,000. Figure 1 shows the probabilities that the real value falls within various ranges around the estimate of 132,000.

[7] Note that this analysis does not cover forced disappearances (they are handled separately). There was insufficient time and resources to extend this analysis to disappearances.

Figure 1: Probability that the actual number of deaths falls within the indicated interval



The number of victims based on the three independent databases

Analysis of Overlap

The information in the three databases lists victims identified or estimated by witnesses. Some, but not all, of the victims were identified by name.[8] The total number of killings in each database is referred to using the notation described below.

M_{CEH} = the total number of victims in the CEH database
M_{CIIDH} = the total number of victims in the CIIDH database
M_{REMHI} = the total number of victims in the REMHI database

None of the three databases directly estimates the total number of killings in the country during the full period of the CEH mandate. Each database is a list of victims of killings who were reported directly to the project and verified according to the methodology of that particular project. As has been mentioned, many victims were not reported to any project. The total number of victims killed in Guatemala and reported (or not) to different projects can be expressed by eight categories, defined below.

N_{000} = victims who were not reported to any of the three projects: not to the CEH, the CIIDH, nor to REMHI
N_{111} = victims who were reported to all three projects
N_{110} = victims reported to the CEH and to the CIIDH, but not to REMHI
N_{101} = victims reported to the CEH and to REMHI, but not to the CIIDH
N_{011} = victims reported to the CIIDH and to REMHI, but not to the CEH
N_{100} = victims reported only to the CEH, and not to the CIIDH nor to REMHI
N_{010} = victims reported only to the CIIDH, and not to the CEH nor to REMHI
N_{001} = victims reported only to REMHI, and not to CEH nor to the CIIDH

The total number of victims of killing in Guatemala, N, is the sum of these eight values. The total number of victims reported to one, two, or three projects, N_k, is the sum of the seven categories that are calculated directly from the databases, that is, N_{111} to N_{001} as shown in Equation 1, below.

[8] This analysis treats victims, not violations, but for killings, the two measures are identical and so this distinction is not significant. See Ball (1996).

$$\hat{N}_k = N_{111} + N_{110} + N_{101} + N_{011} + N_{100} + N_{010} + N_{001} \quad (1)$$

To get the total number of victims reported to one or more projects, estimates must be made of the number of victims reported to all the projects (N_{111}, cases that are found in all three database), and those reported to each pair of projects (N_{110}, N_{101}, *and* N_{011}), and the complements of the number reported to each project (N_{100}, N_{010}, *and* N_{001}). With this information, we can determine N_k.

Matching

It is difficult to find the same victim in any two or all three of these databases using a computer program. Victims are reported with varying information. Identical names may be spelled differently, sometimes because they were inconsistently or idiosyncratically translated from Mayan languages. Dates of birth and death can be uncertain or wrong.

Thus, it is neither practical nor accurate to match databases by automated means with computer programs. To find a person from one database (the source) in another of the databases (the target), an analyst must compare all of the data relevant to the killing, including the name, place, and date of the killing from the source with all the records in the target. This process we call "matching."[9]

Many victims are not identified by name in the databases. Often the original witnesses would mention only a group of people. Different witnesses of the same event often estimate different numbers of victims who suffered the same violations. In our analysis we assume that the match rates between unnamed victims are the same as the rates among the identified victims.

Matching databases is difficult, tedious and time-consuming. Instead of trying to match all the records of each database against **all** the records in the other databases, stratified random samples of the victims identified by name in each database were selected and matched against the records in the other two. The samples were proportionally stratified by region to assure that all regions were covered. The number of records taken in each sample is denoted by the letter *m* (m_{CEH}, m_{CIIDH}, m_{REMHI}). Including all the regions, the total number of records sampled and matched was 1,412, 1,351, and 1,122, respectively (see Table 2).[10]

Each person sampled (from each of the three databases) was compared against the records in the other two databases. When the same person was found in one of the other two databases, it was noted as a double-match; when the record was found in both of the other two databases, it was noted as a triple match.

Four groups of samples were chosen from the three databases. One analyst from the CEH matched one group, and a second analyst matched the other three groups. Many records were deliberately included in both samples. Only in a small number of cases were differences between the analysts' decisions found. The implication from this finding is that the error resulting from non-sampling factors was minimal.[11] The numbers of matches are shown in Table 2.

[9] Furthermore, many victims are not identified by name in the databases. When witnesses mentioned a group of victims without specifying the victims' names, different witnesses often refer to different numbers of victims. Given the already-mentioned difficulty that witnesses often confuse the exact dates of the events, it is not possible to match groups of unnamed victims. This analysis assumes that the match rates between unnamed victims are the same as the rates between named victims.

[10] Of the records mentioned in the text, 498 were resampled and matched a second time. We refer to these records in the analysis of the reliability of the matching.

[11] In the match analysis, what concerns us is that records that are true matches do not escape the analysts. Of the 498 records matched twice, 171 were true matches. Comparing these 171 records matched by two different analysts, 88% were coded identically.

Table 2: Number of matched records in three databases, by outcome

	CEH	CIIDH	REMHI
m_{111}	21	73	19
m_{110}	48	153	
m_{101}	210		226
m_{011}		121	27
m_{100}	1,133		
m_{010}		1,004	
m_{001}			850
Sample Total	1,412	1,351	1,122

Table 2 shows that of the sample of 1,412 victims selected from the CEH database, 21 were found in the CIIDH database and in the REMHI database; these 21 are triple matches. Forty-eight victims in the CEH database were found in the CIIDH database but not in the REMHI database. In addition, 210 more victims in the CEH database were found in the REMHI but not in the CIIDH database. A total of 1,133 of the victims sampled from the CEH database were not found in either of the other two databases. The interpretation of the other two columns is the same.

We obtained overlap rates shown in Table 3 by dividing each m_{xyz} in Table 2 by the total number of victims sampled in each database.

Table 3: Overlap rates for three databases

	CEH	CIIDH	REMHI
r_{111}	1.5%	5.4%	1.7%
r_{110}	3.4%	11.3%	
r_{101}	14.9%		20.1%
r_{011}		9.0%	2.4%
r_{100}	80.2%		
r_{010}		74.3%	
r_{001}			75.8%

To interpret this table, note that r_{110} on the second line indicates that 3.4% of the victims in the CEH database are also in the CIIDH database. The database of the CIIDH is smaller than the CEH database; the same estimation from the point of view of the CIIDH is that 11.3% of the victims recorded in the CIIDH database are also in the CEH database.

Note that the differences in the estimations of the rates are not exactly in proportion to the differences in size among the databases. The differences occur because of the variability that results in the process of taking a random sample, and from the errors in matching. We treat these issues in the later section on the analysis of error.

Estimation of the total number of reported victims

As discussed in "Analysis of overlap" (above), the total number of victims of killing was estimated by the sum of seven components as defined in Equation 1, repeated below.

$$\hat{N}_k = N_{111} + N_{110} + N_{101} + N_{011} + N_{100} + N_{010} + N_{001} \quad \text{Equation 1}$$

With the rates from Table 3 and the last line of Table 1 (the total number of victims 1978-1996), the components of N_k based on information in the three databases can be calculated. The results are in Table 4, below.

Table 4: Number of killings, estimated by category and by project

	CEH	CIIDH	REMHI
N_{111}	364	459	357
N_{110}	833	963	
N_{101}	3,645		4,242
N_{011}		761	507
N_{100}	19,663		
N_{010}		6,317	
N_{001}			15,955

However, to calculate N_k, the number of victims common to all three databases, the several estimates of the number of matched records (N_{111}, $N_{110,}$ $N_{101,}$ and N_{011}) must be reconciled. We used the average of each of these four components, providing the totals shown in Table 5, below.

Table 5: Estimated number of killings in all three databases (CEH, CIIDH, and REMHI).

	Mean
N_{111}	393
N_{110}	898
N_{101}	3,943
N_{011}	634
N_{100}	19,663
N_{010}	6317
N_{001}	15,955
N_k	47,803

Thus, our estimate of the unduplicated number of reported killings in the three databases is 47,803. However, we show below, this number is subject to a number of controllable biases.

Demographic Theory and The Estimation of the Total Number of Undocumented Killings

In the previous section, we explained how we estimated the total number of documented killings. In this section, we discuss how to estimate the total number of undocumented killings, denoted N_{000}. Below, in Figures 2a, 2b, and 2c, we show with the use of Venn diagrams, the three possible ways in which the databases might be related.

Figure 2a: Distribution of databases in the universe of all violations (where there is no relation)

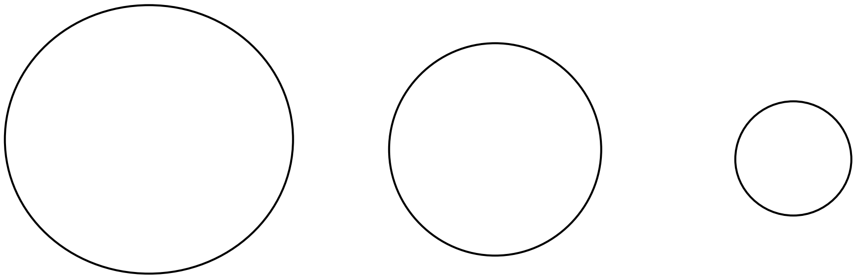


Figure 2b: Distribution of databases in the universe of all violations (total overlap)

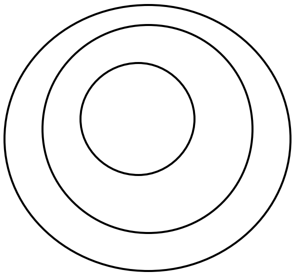
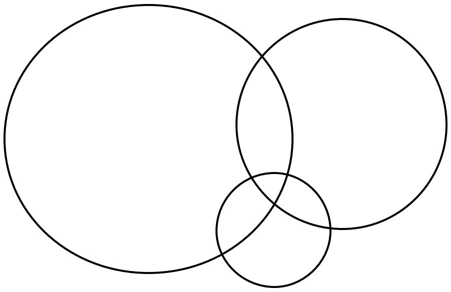


Figure 2c: Distribution of databases in the universe of all violations (partial overlap)



In Figure 2a, the databases share no violations. In Figure 2b, all the violations are contained in the largest of the three databases. In Figure 2c, some violations are shared. From the previous section, it is clear that Figure 2c is the correct representation of the three databases.

Assume that the three projects operated independently and consequently, that the probability that a project has testimony about a certain violation has no influence on whether another project has testimony about the same violation. What implication does this have for the universe of violations? In Figure 2a, the implication is that the universe of violations is large because working independently, the databases do not overlap. In Figure 2b, the implication is the inverse, that the one database is contained within the next larger, and the next larger is contained within the largest. In Figure 2c, which corresponds to our situation, the levels of overlap are partial. With the assumption of independence and the reality of overlap, the number of violations in the universe can be inferred.

Consider the case of two projects, P_A and P_B, whose databases have an overlap M in a universe of violations N.[12] Note that the probability of any given killing being documented by Project P_A is $\Pr(A)=\frac{A}{N}$ that is $N=\frac{A}{\Pr(A)}$, and the probability of any given killing being documented by Project P_B is $\Pr(B)=\frac{B}{N}$. The probability that a killing was documented by both databases, $\Pr(M)$, is equal to $\Pr(M) = M/N$, and by the definition of an event composed of two independent events, $\Pr(M) = \Pr(A|B) = \Pr(A) * \Pr(B)$.

Interchanging the terms, $\Pr(A)=\frac{\Pr(M)}{\Pr(B)}$, which reduces to $\Pr(A)=\frac{M/N}{B/N}=\frac{M}{B}$

Combining the first relation $\Pr(A)=\frac{A}{N}$ with the previous result gives us $\frac{A}{N}=\frac{M}{B}$, and therefore $N=\frac{AB}{M}$. In order to estimate only the killings that were excluded from the two projects, $N_{00}=\frac{(A-M)(B-M)}{M}$, or in the notation of the three-database system,

$$N_{00}=\frac{N_{10} * N_{01}}{N_{11}} \tag{2}$$

With the same logic, it is possible to derive an estimator for $n_{000:}$ the measure of the number of killings that were not documented by any of the three projects.[13] This estimator is presented below in Equation 3.

$$N_{000}=\frac{N_{100}N_{010}+N_{100}N_{001}+N_{010}N_{001}}{N_{110}+N_{101}+N_{011}} \tag{3}$$

Measuring the Sampling Error and the Estimator Error

The estimators of N_{000} and of the total number of killings $\hat{N}$ depend on the estimates of the overlap between the three databases.[14] To estimate of the number of killings in the categories N_{111}, N_{110},…, N_{001} that sum to $N_{k,}$ the levels presented in Table 3 above are multiplied by the total number of killings in each database. We then estimated n_{000} using Equation 3, above. Summing these two components gives an estimate of $\hat{N}$. We used the jackknife method to estimate $\hat{N}$ because that method allows us to control the ratio bias inherent in n_{000} and to estimate the variation in the three estimators necessary for this analysis (N_k, N_{000}, and $\hat{N}$). In the general explanation of the method below, the estimator $\hat{q}$ represents each of the three estimators. For example, N_k in Table 5, above, (47,803), is $\hat{q}$ for N_k taken at the national level.

The method first randomly divides the sample of matched records (total size n records) into k groups, each of which contains m records such that $n = mk$. $\hat{q}_{(a)}$ is calculated by the same method as $\hat{q}$ but with the sample reduced $m(k-1)$ obtained by omitting group a.

Define

$$\hat{q}_a = k\hat{q} - (k-1)\hat{q}_{(a)} \tag{4}$$

and

$$\hat{\hat{q}} = \frac{1}{k}\sum_{a=1}^{k} \hat{q}_a \tag{5}$$

Equation 4 gives us k values of $\hat{q}_a$ calculated from the sub-samples reduced by omitting a group k; the average of these values is $\hat{\hat{q}}$ (see equation 5), called Quenouille's estimator, and removes various biases that affect $\hat{q}$. This estimator is presented in Table 7.

[12] This explanation is taken from Marks, Seltzer, and Krótki (1974, pp. 13-17).

[13] See (Marks, Seltzer, and Krótki, 1974, equation 7.188) . Two possible estimators are given, but we chose the one preferred in cases such as ours, where there is likely to be correlation bias.

[14] This section is largely based on Wolter (1985, pp. 154-155).

The other beneficial result of the jackknife method is that the values of $\hat{q}_a$ are distributed approximately normally.[15] The standard error of the estimator (the square root of the variance) is estimated in Equation 6.

$$SE(\hat{\bar{q}}) = \sqrt{\frac{1}{k(k-1)}\sum_{a=1}^{k} (\hat{q}_a - \hat{\bar{q}})^2} \qquad (6)$$

The standard errors given in Table 7 were calculated with equation 6.

Coverage in Space and Time and its Effect on *N₀₀₀*

In the discussion concerning Table 1 we noted that none of the three projects covered well violations in the period 1960-1977, and thus no estimation of $\hat{N}$ for this period is possible. The most important complication for the estimation of N_{000} is that the projects did not cover all of the regions with the same intensity. If the regions with different coverage rates among the three projects are not handled separately, the estimation could be biased.

For example, consider Region IV, in which the CEH found more than 1,300 killings, while the other two projects reported only some dozens of killings. Clearly the levels of overlap are low, but the overlap rates should not be used for an estimate of N_{000} because the concept of overlap requires that the projects actually collected data in the same areas. Therefore the estimation of n_{000} was based only in the projects that were in fact able to work in each region. The projects that contributed to the estimation of n_{000} are in Table 6.

[15] The "pseudovalues" $\hat{q}_a$ should be approximately independent and distributed identically. This assumption was tested with a normal probability plot for each set of pseudovalues, and in each case the results were consistent with this assumption.

Table 6: Projects used to estimate N_{000}, by region

Region	Projects with adequate coverage	Equation for n_{000}
Region 0 (other)	All three	Equation 3
Region I (área Ixil)	All three	Equation 3
Region II (Cahabón)	All three	Equation 3
Region III (Rabinal)	All three	Equation 3
Region IV (San Martín Jilotepeque)	Only CEH	Unable to estimate n_{000}
Region V (North of Huehuetenango)	CEH & REMHI	Equation 2
Region VI (area Zacualpa)	All three	Equation 3
Region VII (Guatemala)	CEH & REMHI	Equation 2
Region VIII (area Panzós)	CEH & REMHI	Equation 2
Region IX (Ixcán)	CEH & REMHI	Equation 2
Region X (area Costa Sur)	All three	Equation 3

Results and Some Limits on the Interpretation

We now show the results by component and by region.

Table 7: Total number of killings in Guatemala 1978-1996, by category of estimation and region

	Region											
Category	**0**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**Total**
N_{111}	67	141	15	146	0	0	17	2	0	2	2	391
N_{110}	378	406	8	98	5	0	67	3	0	16	2	983
N_{101}	1,358	1,010	204	170	13	206	336	24	43	681	13	4059
N_{011}	133	419	16	122	0	0	0	0	0	0	0	690
N_{100}	8,260	3,187	221	1,028	1,325	1,597	1,642	226	156	1,720	182	19,545
N_{010}	2,256	2,708	85	836	16	1	195	15	41	30	91	6,274
N_{001}	5,228	3,999	295	926	59	765	1,166	77	1,099	2,054	106	15,773
N_k (without duplication)	17,679	11,870	844	3,328	1,418	2,569	3,416	347	1,339	4,501	396	**47,706**
SE (N_k)	110	135	24	79	11	44	76	9	26	77	5	228
N_{000}	38,856	17,397	466	6,467	0	5,548	5,836	561	2,265	5,052	2,019	**84,468**
SE (N_{000})	3,809	2045	105	1,152	0	1,826	1,890	350	3,062	995	1,840	6,388
$\hat{N}$	56,535	29,267	1,310	9,795	1,418	8,117	9,252	908	3,604	9,553	2,415	**132,174**
$SE(\hat{N})$	3,918	2175	127	1,218	11	1,870	1,964	357	3,087	1,072	1,844	6,568

In Table 7, it can be seen that in Region 0 there were N_k = 17,679 killings documented in all three projects. Over all regions, there were 47,706 killings documented, being the sum of the regional estimations[16]. The standard error SE(N_k) is not the simple sum of the regions, but rather the square root of the sum of the squares of the regional values (i=0, I, …, X):

$$SE(N_{kl}) = \sqrt{\sum_{i=0}^{X} SE(N_{ki})^2} \qquad \text{Equation 7}$$

Similarly, the values for N_{000} and for N are the sum of the regional values and the standard error of N_{000} and $\hat{N}$ is the square root of the sum of the squared regional values. In this way, we estimate

[16] The estimation for N_k was 47,706 murders documented between all three projects, with a standard error of 228, yielding a 95% confidence interval of 47,559 to 48,152. Note that this range includes the value estimated in Table 5, 47,803. The closeness of the value in Table 5 with the value estimated by the sum of the regions by the jackknife method implies that there is not much bias in the simple estimation. Nonetheless, the bias that required the disaggregation by regions may not have affected N_k, but yet might still affect N_{000}, and is therefore still necessary.

that there were approximately 84,468 killings that were not reported to the CEH, to the CIIDH, or to the REMHI project. Summing N_k and n_{000} to $\hat{N}$, we have as our final estimate, that there were 132,174 killings in Guatemala between 1978-1996, with a standard error of 6,568.

Possible Corrections and Limitations to the Interpretation of Table 7

There are five sources of error that cannot be quantified in this analysis due to lack of time, resources, and adequate data. In preliminary analysis of this type, the global effect of these corrections is conservative, in that the corrections tend to reduce the estimation of $\hat{N}$. The conclusion of this section is that the accumulated effect of the identified biases is does not significantly change our interpretation of Table 7.

Correlation between sources

The estimation of N_{000} depends on the assumption of independence between sources; that is, that the probability that any given respondent gives her testimony to one project has no effect on the probability that the same respondent will give her testimony to one of the other two projects. It is certain that this correlation is not zero, but is positive, for two reasons.

First, psychological research has shown that survivors of human rights violations who are able to give testimonies under supportive conditions experience improvement in their mental health. Thus, it is likely that people who give testimony in these conditions may seek additional opportunities to give their testimony, thereby increasing the overlap rate.

Second, it is known that several popular movement groups organized their social bases to present testimonies to all three projects. In this way, members of these organizations would have greater probabilities of giving testimonies more than once, thus, reporting the same violations more than once, and increasing the overlap levels. The two effects – both of which are certain – would tend to bias the estimate of n_{000} toward a smaller number.

Matching errors

If the analysts who conducted the matching failed to find victims who were in multiple databases, by accident or because there were inadequate data in the original sources, these omissions would tend to depress the level of measured overlap and in consequence bias the estimation of n_{000} upwards. In preliminary investigations, (all that are possible given the partial state of many cases) only minimal effects of this kind were found. At most, they amounted to about 12% of the final estimate of $n_{000,}$ implying about 8% of $\hat{N}$. Considering the other sources of error listed in this section, and recognizing that the data for this analysis were limited, we decided not to quantify this error (or a correction for it) in the final analysis.

Internal duplication

All of the projects that receive information from primary sources may have problems with internal duplication that results from multiple reports of the same events.[17] Internal duplication tends to artificially increase the number of killings that are reported in a single database. All three projects worked hard to clean their data to reduce internal duplications, but some always remain. In a preliminary analysis, insufficient duplication appeared to require a correction for this source of error.

Rates of overlap between kinds of victims

The amount of overlap between the three databases was based on an analysis of victims identified by name and surname. However, many victims are not identified by name as a result of large-scale violence that overwhelmed the capacity of the witnesses to remember all the victims' names. It is possible that the level of overlap between victims not identified by name could be either higher or lower than the levels measured among named victims. Given the difficulty of matching unnamed victims, it is not possible to quantify the potentially different overlap levels among victims with

[17] See, in this context REMHI (1998, pp. XXXI-XXXII), and Ball, Kobrak, and Spirer, (1999, pp. 62, note 12).

different amounts of identifying information. As mentioned earlier, we assumed that the match rates for unnamed and named victims were the same.

Geographic areas excluded from the analysis

In Table 7 we noted that there is not sufficient data to estimate n_{000} in Region IV. Given the experience of other regions, in which the ratio between N_k and n_{000} varies between 0.5 and 2, with a mode and mean close to 0.5, it is likely that the value of n_{000} for Region IV is approximately 2,500. In other regions N_k is composed of data from more than two projects. However, in Region IV information was collected only by the CEH. Thus, it is possible that the ratio between N_k and n_{000} for Region IV could be 0.25 or less, increasing the estimate by a factor of two or more. There are no other methods available to reduce the lack of certainty about this number, and therefore it is not included in the final estimate.

In addition, Region IV exemplifies a more fundamental problem: This methodology works only for areas covered by at least two of the three projects, even if the two projects only partly cover each area. In areas in which only one or none of the three projects conducted interviews, there is no basis for an estimate of the total number of excluded victims (n_{000}). Instead, in these situations only N_k (the total documented number of killings) enters the estimation process. As all three projects focussed on particular areas in Guatemala in which large scale violations were known to have occurred on the basis of journalistic and NGO accounts, it is unlikely that the excluded areas had high levels of human rights violations. However, if our analysis included any excluded area in which killings could have occurred, that inclusion would tend to increase the final estimate.

Estimation of Killing Rates, by Ethnic Group and Region

The rate of killing for a defined group is the proportion of people in that group who were killed. Quite simply, it is the number of people in the group who were killed divided by the total number in the group prior to the killings. The CEH was interested in comparing the rates of killing for defined ethnic groups during the period 1981-1983.

Six geographical regions were identified as those in which -- according to secondary sources and anecdotal evidence -- state violence was concentrated against indigenous peoples. These six regions are listed in Table 8, with the ethnic group populations according to the census of 1981.

Table 8: Populations in six regions, by ethnic group, 1981

	Indigenous	Non-indigenous
Region I: Ixil area	38,902	5,882
Region II: Cahabón	20,706	868
Region III: Rabinal	18,610	4,120
Region IV: San Martín Jilotepeque	31,690	4,876
Region V: north of Huehuetenango	53,556	11,123
Region VI: Chiché, Zacualpa, Joyabaj	51,105	10,997

To calculate the killing rate, the number of victims is estimated. We did this estimation twice, first to get the total number of documented victims, and then to get the estimated number of victims using the methods outlined above.

The following are the steps in this estimation process:

1. The number of murders that occurred 1981-1983, less those attributed to the URNG, were calculated by the ethnic group classifications indigenous, not indigenous, and unknown, for each of the six regions in the three databases. This step is analogous to Table 1 above,

but limited to killings attributed to the state during the period 1981-1983 and disaggregated by the ethnicity of the victims.

2. The number of matched victims and the corresponding rates were calculated for each of the six regions (logically following the method shown for Tables 2 and 3).[18]
3. The number of victims for each ethnic group was estimated using the regional rates of overlap and the number of victims in each database (similar to Table 4).
4. The estimates were made for each region by taking the average of each of the three database estimates (similar to Table 5).
5. The jackknife method was applied to each defined group, following equations 4, 5, and 6, in order to estimate N_k and $\hat{N}$ (and their standard errors) for each ethnicity in each region. The values of N_k are presented below in Tables 9a and 9b, and those for $\hat{N}$ in Tables 11a and 11b.
6. Taking from Table 9a the victims with known ethnicity, the victims without known ethnicity were apportioned to the categories “indigenous” or “not-indigenous” according to the proportions shown below in Table 10, creating the figures shown in Table 9b.
7. With the information from Tables 8, 9, and 10, the proportion killed of each ethnic group in each region was calculated, along with its standard error. The data, presented below in Figure 3, explain *inter alia* that according to the information documented by the CEH, CIIDH, and REMHI, more than 14% of the indigenous population in the Ixil area in 1981 were murdered by 1983, while in the same period and area, only 2% of the non-indigenous population were killed.

[18] The overlap rates were not calculated by ethnic group. Instead the regional match rates for the period 1981-1983 were applied equally to the ethnic groups in that region. This application assumes that the overlap rates did not vary significantly among ethnic groups.

Table 9a: Number of documented killings (N_k) in three databases, 1981-1983, by region and ethnicity

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	N_k	1388	340	1071	1012	1020	1126
	$SE(N_k)$	25.5	13.12	33.71	0.67	17.11	16.7
Non-indigenous	N_k	32	2	13	16	8	6
	$SE(N_k)$	0.49	0.07	0.33	0.14	0.12	0.13
Unknown ethnicity	N_k	4339	186	1669	10	752	1208
	$SE(N_k)$	62.87	9.02	47.75	0.64	31.95	46.79

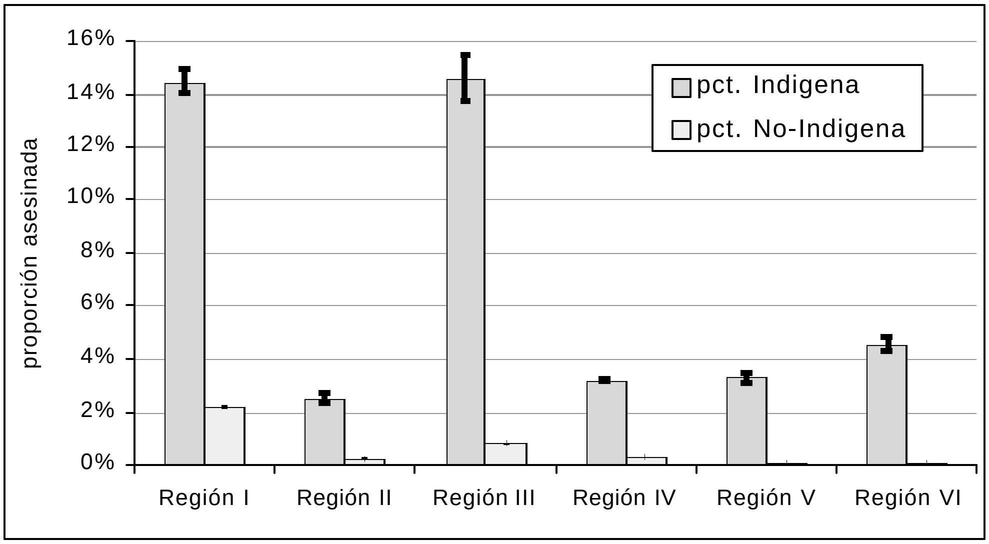
Table 9b: The number of documented killings (N_k) in three databases, by ethnic group, including victims without known ethnicity, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	N_k	5,632	525	2,720	1,022	1,767	2,327
	$SE(N_k)$	66.56	15.90	57.98	0.92	36.05	49.42
Non-indigenous	N_k	127	3	33	16	13	13
	$SE(N_k)$	0.49	0.07	0.33	0.14	0.12	0.13

Table 10: Percentage of indigenous of victims with known ethnicity of all victims in Table 9a

	Region I	Region II	Region III	Region IV	Region V	Region VI
Proportion indigenous	97.8%	99.6%	98.8%	98.5%	99.3%	99.4%
Proportion of victims with known ethnicity	24.7%	64.7%	39.4%	99.1%	57.7%	48.4%

Figure 3: Documented proportion of the population killed by State forces in Guatemala 1981-1983, by region and ethnic group, with the 95% confidence interval[19]



8. Note that the data presented for Region VI (Zacualpa area) in Figure 3 do not correspond exactly to the statistics presented in the genocide section of the CEH report because the definition of Region VI used here includes the *municipios* of Chiché, Joyabaj and Zacualpa. In the genocide section of the report, only the *municipio* of Zacualpa is considered part of Region VI. The statistics for Region VI (and for all the regions) in Figure 3 and in the genocide section were calculated with the same methods but with different population and violation bases.

9. The projected totals ($\hat{N}$) by ethnicity and region were calculated using the same methods described with equations 2-6, and with the same data as shown in Tables 8-10. The statistics are presented in Tables 11a and 11b and rates are shown in Figure 4.

[19] Source of the graph: 1981 census; testimonies received by the CEH, direct sources to the CIIDH, and testimonies received by the REMHI project.

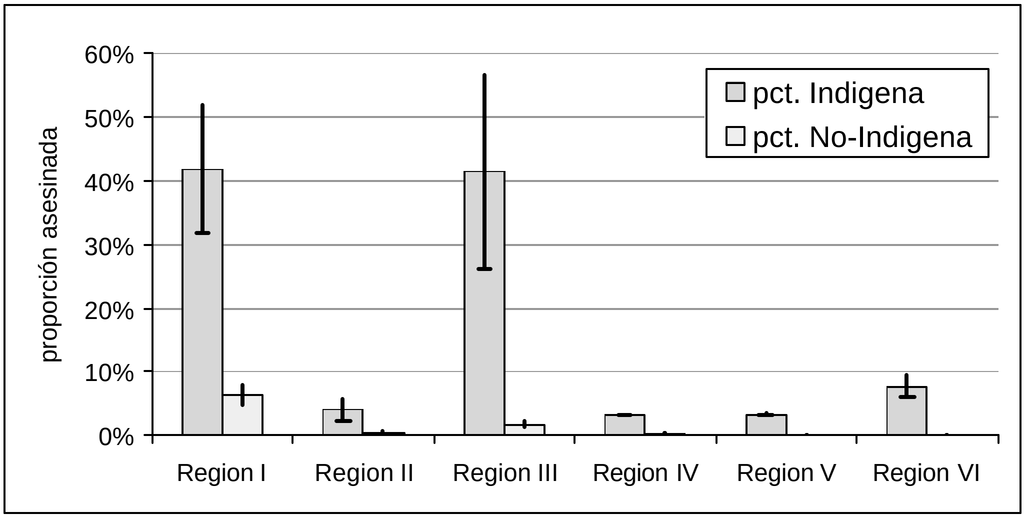
Table 11a: Number of projected killings ($\hat{N}$) in three databases, by ethnic group and region, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	$\hat{N}$	2578	443	2983	1012	1020	2723
	SE($\hat{N}$)	190.28	39.13	587.58	0.67	17.11	386.43
Non-indigenous	$\hat{N}$	63	2	13	15.7	7.5	6
	SE($\hat{N}$)	5.4	0.07	0.33	0.14	0.12	0.13
Unknown ethnicity	$\hat{N}$	14014.6	394	4791	10	752	1208
	SE($\hat{N}$)	1841.44	151.12	874.1	0.64	31.95	46.79

Table 11b: Number of projected killings ($\hat{N}$) in three database, by ethnic group and region, including victims without identified ethnicity, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	$\hat{N}$	16,284	835	7,717	1,022	1,767	3,924
	SE($\hat{N}$)	1,811.0	155.5	1,044.5	0.9	36.1	389.2
Non-indigenous	$\hat{N}$	371	4	70	16	13	13
	SE($\hat{N}$)	40.9	.6	10.5	.1	.3	.3

Figure 4: Projected proportions of ethnic groups killed by state forces in Guatemala 1981-1983, by region and ethnic group, with 95% confidence interval



Note that Figures 3 and 4 have two interpretations.[20] First, regions I and III were the most affected by state violence (based on rates). In these two regions there are clear quantitative signs that the killing was so massive that it could have been genocide. Second: in all the regions the victims were disproportionately indigenous. Note, for example that as shown in Figure 4, in Region I more than 40% of the indigenous population was killed while approximately 8% of the non-indigenous population were killed. The difference between the killing rates is a factor of five. In the structure of violence committed by the Guatemalan state, these are revealing differences.

Comparisons among Databases

In the analysis of multiple databases, the databases can be compared in order to determine the levels of temporal or geographic agreement among them, and in terms of the relative levels to which they attribute responsibility to the state or insurgent forces. In this section we compare the tendencies and statistics among the three databases.

It is clear from Table 1, that there is only a moderate level of agreement about where violence occurred among the databases. In some regions all three projects found many violations, (Regions 0, I, and III), while in other regions only two projects investigated deeply (for example, Regions V, VI, IX), while in Region IX only the CEH carried out an intensive investigation.

Although the projects covered distinct areas, the second section in this part shows that the databases agree on which months saw the peaks of the violence. The third section considers the relative proportions of responsibility attributed to the two parties to the conflict.

Coincidence in Time

If the months are ordered in terms of how many killings are reported according to each of the three databases, a relatively high level of agreement is found. In Table 12, the top ten months are shown ordered as described, presenting the percentages of the total number of killings during the entire period 1979-1984.

Table 12: The ten most violent months in the three databases, 1979-1984[21]

	CEH				CIIDH				REMHI		
Range	**Month**	**Total**	**Pct.**		**Month**	**Total**	**Pct.**		**Month**	**Total**	**Pct.**
1	82-01	2,256	9%		82-02	610	12%		82-03	1,330	12%
2	82-03	2,253	9%		81-06	390	7%		82-02	807	7%
3	82-02	1,880	8%		83-03	297	6%		82-07	792	7%
4	82-08	1,819	8%		82-06	279	5%		82-05	657	6%
5	82-07	1,719	7%		82-07	234	4%		81-09	629	6%
6	81-01	1,423	6%		82-01	233	4%		82-01	470	4%
7	82-06	1,146	5%		82-04	222	4%		82-04	428	4%
8	82-04	937	4%		82-05	210	4%		81-07	397	4%
9	82-05	895	4%		83-08	180	3%		80-02	364	3%
10	81-09	754	3%		81-02	174	3%		82-10	360	3%
Ten month total		**15,082**	**63%**			**2,829**	**54%**			**6,234**	**56%**
Total for 1979-1984		**23,890**	**100%**			**5,275**	**100%**			**11,065**	**100%**

The shaded five months are those for which the three databases show concordance. Within the ten worst months in each database, the three systems agree on five months: January, February,

[20] Note that in both absolute and relative terms, the standard error for each statistic in Figure 3 is greater than that for the analogous statistic in Figure 2. This is consistent, as the projections in Figure 3 incorporate more uncertainty than the estimations in Figure 2. The size of the samples on which the estimation of the unduplicated totals were based (N_k) are sufficient for those estimations which do not have such high errors as to make them unusable. The projection, however, still contains significant uncertainty, reflected in the higher error rates.

[21] In Table 12, only killings identified with dates precise to the month are included.

April, May, and July of 1982; other months of the same year (March and June) coincide in two of the three databases.

The databases are in agreement that approximately half of the killings occurred in the ten worst months (63%, 54%, and 56%). This concentration follows Pareto's Law, which states that 80% of any given phenomenon will occur in 20% of the categories. However, the closeness of these months to each other in time (all of them occur toward the first half of 1982) is strong evidence that this period is the most intense period of political violence. Furthermore, the level of agreement between the databases implies that although the projects did not investigate exactly the same regions of the country, they found the same trends in time.

Coincidence in the Attribution of Responsibility

The three sources agree in the attribution of responsibility to state forces and the insurgents: together they attribute more than 94% of the killings to the state and less than 6% to the guerrilla forces.

Table 13a: Total number of killings with identified perpetrator, by responsible entity[22]

	CEH	CIIDH	REMHI	TOTAL
State	24,121	2,916	19,177	46,214
Guerrilla	1,263	61	1,184	2,508

Table 13b: Percentage of killings with identified perpetrator, by responsible entity

	CEH	CIIDH	REMHI	TOTAL
State	95%	98%	94%	95%
Guerrilla	5%	2%	6%	5%

In an analysis of the attribution of responsibility derived from non-probability samples, such as those used for the statistics in Tables 13a and 13b, there is the possibility that the data may have been biased towards the violations committed by one entity or another because of a tendency to focus on one perpetrator. This kind of "over-focus" bias can effect the estimated proportions.

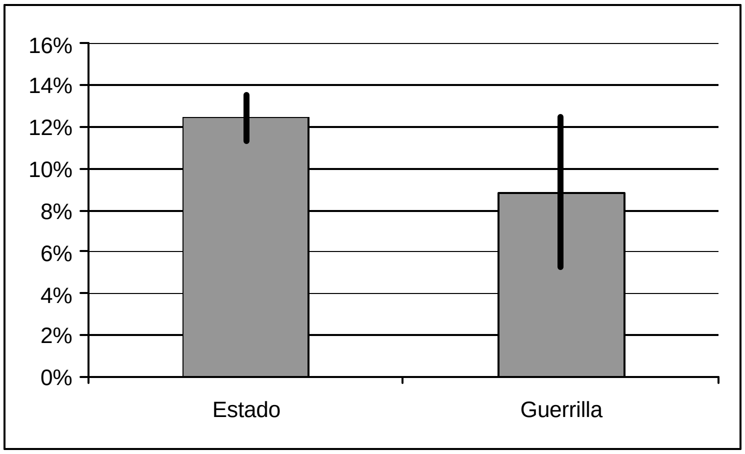
For example, if the projects looked primarily for violations committed by the guerrilla forces, while dedicating less effort to the search for violations committed by state forces, this data would reflect an inflated level of responsibility attributed to the guerrilla forces. In a probability sample, this bias is avoided by accepting testimonies that were selected at random. None of the three projects sought testimonies according to a probabilistic design and hence, there may be bias in these resulting proportions.

Taking advantage of the existence of the three projects and using the measures of overlap explained earlier, we can test the hypothesis that the projects over-focused on one or the other of the perpetrating entities. The components of the estimated documented killings must be separated into the killings attributed to the insurgents and those attributed to the state. The components that indicate overlap are summed ($N_{111}+N_{110}+N_{101}+N_{011}$) and divided by N_k; this figure gives the percent of overlap by responsible group. The results of this calculation are shown below in Figure 5[23].

[22] This only includes violations with the perpetrator identified and with a date precise to the year. It is worth reemphasizing that this analysis includes all killings, not only arbitrary executions.

[23] The table is based on Figure 5, corresponding to Table 5 divided by the perpetrating entity, shown in disaggregated form below.

Figure 5: Overlap rates for victims of killing documented by the CEH, the CIIDH, and by REMHI, for violations committed by the state and guerilla forces (with bars to indicate the 95% confidence intervals.).



With the results of Figure 5, we rejected the hypothesis that there is a significant difference between the level of coverage of violations committed by the guerrilla forces and those committed by the state forces. Although there is a small difference in the overlap rates of the state forces (12.4%) and the guerrilla forces (8.8%), the difference is within the standard error. Thus, the difference cannot be distinguished from the sampling error of the matching process.

The difference between the overlap rate for killings committed by the state and by the guerrillas is significant, neither in technical terms, nor in analytic terms. The analysis of the effect on the estimate proportions follows. The technical test is the following:

$$SE = \sqrt{\frac{p_E * (1 - p_E)}{N_E} + \frac{p_G * (1 - p_G)}{N_G}} = 0.0193,$$

which yields the confidence interval +/- 3.8%. The difference between the two rates is 12.4% - 8.8% = 3.6%; the confidence interval is more than the difference, which means that we cannot reject the hypothesis that the difference is equal to zero. This calculation confirms the intuitive interpretation from Figure 5.

The implication of Figure 5 is that all three projects investigated violations committed by the guerrillas and violations committed by the state with approximately the same level of coverage and intensity. Therefore there is no systematic disproportionality in the intensity of investigation between the two entities sufficient to change the interpretation of the proportions of responsibility attributed to each.

Category	State	Guerrilla
N_{111}	299	3
N_{110}	617	
N_{101}	3769	200
N_{011}	388	
N_{100}	19,173	1,087
N_{010}	2,165	61
N_{001}	14,430	949
N_k	40,842	2,301
Overlap Rate $\frac{N_{111} + N_{110} + N_{101} + N_{011}}{N_k}$	12.4%	8.8%
SE (N_k)	0.5%	1.9%

The standard error calculated by the conventional method for proportions derived from samples is $SE = \sqrt{\frac{p * (1 - p)}{N}}$, and the 95% confidence interval is +/- 1.96*SE.

There are two ways to consider the effect of that the overlap rates on the proportion of responsibility attributed to the state and the guerrillas. The proportions of attributed responsibility that result from N_k estimated in note 23, are presented below. The average does not come from all three databases, as implied in Table 13 in the text. Note that this analysis excludes violations for which responsibility is unclear; adding the unknown category would reduce both proportions slightly.

Figure 6. Proportions of attributed responsibility.

Estimation	**State**	**Guerrillas**
N_k	41,147	1,860
Proportion of the total N_k	95.7%	4.3%
n_{000}	73,622	3,706
$\hat{N}$	114,769	5,567
Proportion of the total $\hat{N}$	95.4%	4.6%

To see the insignificant effect of the disproportionality on coverage, n_{000} is calculated (using Equation 3 for the state, but using Equation 2 for the guerrillas because the CIIDH did not report sufficient guerrilla violations). The estimation of n_{000} includes the information about the overlap rates, and in this way n_{000} controls the effect of the disproportionality in coverage. Note that the calculated proportions of $\hat{N}$ are the same as those calculated for N_k. The conclusion is that the disproportionality in coverage of the state and the insurgents does not change the final analysis about their relative responsibility.

Appendix 1

Regional Definitions, by *municipio*

All of the *municipios* other than those listed here were classified as Region 0. Note that the definition of Region VI in this study is not the same as the definition used in the CEH report. This study included two additional *municipios* that the genocide study did not include.

REGION	*Departamento*	*Municipio*
Region I	Quiché	Chajul
Region I	Quiché	San Juan Cotzal
Region I	Quiché	Nebaj
Region II	Alta Verapaz	Cahabón
Region III	Baja Verapaz	Rabinal
Region IV	Chimaltenango	San Martín Jilotepeque
Region V	Huehuetenango	Nenton
Region V	Huehuetenango	San Mateo Ixtatán
Region V	Huehuetenango	Barillas
Region VI	Quiché	Chiche
Region VI	Quiché	Zacualpa
Region VI	Quiché	Joyabaj
Region VII	Guatemala	Guatemala
Region VII	Guatemala	Mixco
Region VIII	Alta Verapaz	Panzós
Region VIII	Alta Verapaz	San Pedro Carchá
Region IX	Quiché	Ixcán
Region X	Santa Rosa	Cuilapa
Region X	Santa Rosa	Barberena
Region X	Santa Rosa	Casillas
Region X	Santa Rosa	Santa Rosa De Lima
Region X	Santa Rosa	Oratorio
Region X	Santa Rosa	San Rafael Las Flores
Region X	Santa Rosa	Santa Maria Ixhuatan
Region X	Santa Rosa	Taxisco
Region X	Santa Rosa	Chiquimulilla
Region X	Santa Rosa	San Juan Tecuaco
Region X	Santa Rosa	Guazacapán
Region X	Santa Rosa	Naranjo
Region X	Santa Rosa	Pueblo Nuevo Las Viñas
Region X	Santa Rosa	Nueva Santa Rosa
Region X	Escuintla	Escuintla
Region X	Escuintla	Santa Lucía Cotzumalguapa
Region X	Escuintla	La Democracia
Region X	Escuintla	Siquinalá
Region X	Escuintla	Masagua
Region X	Escuintla	Tiquisate
Region X	Escuintla	La Gomera
Region X	Escuintla	Guanagazapa
Region X	Escuintla	San José
Region X	Escuintla	Iztapa
Region X	Escuintla	Palín
Region X	Escuintla	San Vicente Pacaya
Region X	Escuintla	Nueva Concepción
Region X	Retalhuleu	Retalhuleu
Region X	Retalhuleu	San Sebastián
Region X	Retalhuleu	Santa Cruz Mulua
Region X	Retalhuleu	San Martín Zapotitlán
Region X	Retalhuleu	San Felipe
Region X	Retalhuleu	San Andrés Villa Seca

REGION	*Departamento*	*Municipio*
Region X	Retalhuleu	Champerico
Region X	Retalhuleu	Nuevo San Carlos
Region X	Retalhuleu	El Asintal
Region X	San Marcos	Nuevo Progreso
Region X	San Marcos	El Tumbador
Region X	San Marcos	Malacatán
Region X	San Marcos	Catarina
Region X	San Marcos	Ayutla
Region X	San Marcos	Ocos
Region X	San Marcos	Pajapita

References

Ball, Patrick, 1996. *Who Did What to Whom? Designing and Implementing a Large-Scale Human Rights Data Project*. Washington: American Association for the Advancement of Science.

Ball, Patrick, Kobrak, Paul, and Spirer, Herbert, 1999. *State Violence in Guatemala, 1960-1996: A Quantitative Reflection*. Washington: American Association for the Advancement of Science and Centro Internacional por Investigaciones en Derechos Humanos.

Marks, Eli S., Seltzer, William, and Krótki, Karol J., 1974. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.

REMHI 1998. Project Report, *Guatemala: Nunca Más Tomo IV, Victims del Conflicto*. Guatemala: Oficina de Derechos Humanos del Arzobispado de Guatemala.

Wolter, Kirk M., 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag.