Skeptical About the Reasoning of the Bible Code Skeptic

Robert M Haralick
Computer Science
Graduate Center
City University of New York
New York, NY 10016

In the November/December 1997 issue of *Skeptical Inquirer*, physicist David Thomas wrote a critical article *Hidden Messages and the Bible Code.* The article shows Bible code tables of related key word pairs such as Roswell with UFO and Nazi with Hitler in English texts which everyone agrees has no hidden messages. The examples illustrate the conventional wisdom that tables can be found in any text. That being so, their inference is that the claim that the code tables found in the Hebrew text of the five books of Moses are unusual is most assuredly false.

However, on a careful reading of the article, it becomes apparent that one may be skeptical of its reasoning. In this note we explore how one would use sound reasoning and statistical inference to determine whether or not there are hidden messages in any given text, pointing out the errors in the examples of the Thomas article as well as some similar examples on McKay's website. We assume the reader is familiar with the basic concepts.

Unusualness must be evaluated in terms of probability. A table that is unusual is one being relatively compact. This that the probability of a table as compact occurring as a chance phenomena is sufficiently small. To estimate a probability in this context requires an experiment. An experiment has a protocol. In the case of the Bible codes, the protocol must consist of ten parts:

(1) the given text to be tested
(2) how a priori sets of key words are determined
(3) specification of how the min and max skip intervals for ELSs of each key word are determined;
(4) the criteria determining for each ELS what code cylinder sizes are searched over;
(5) a control text population all of whose texts must satisfy the hypothesis being tested.
(6) how many texts are to be randomly sampled from the control text population;
(7) the compactness definition which measures closeness between ELSs;
(8) the hypothesis being tested and against what alternative hypothesis it is being tested;
(9) the significance level of the hypothesis test;
(10) the statistical analysis methodology by which the p-value of the hypothesis test will be determined from the experimental data.

In our case the hypothesis being tested is the Null hypothesis of No Bible Code effect. This means that whatever tables one finds in the given text is just a chance occurrence.
The table compactness value indicates that the table is not relatively compact. The alternative hypothesis might be *key words of logically/historically related events have ELSs tending to form more compact formations in the given text than expected by chance.* The p-value of the hypothesis test is the probability that under the Null hypothesis a result as good or better than the one observed from the given text would be obtained from a randomly selected text from the the control text population. In essence it is the fraction of the texts in the control text population that have better tables than the ones found in the given text.

In their examples, what Thomas and McKay have done is designate the text control population to be novels. They select one novel from the control text population. They find a table for some pair of key words (not necessarily a priori). They show us the table. The logic of their demonstration goes something like this: if what is claimed to be found as unusual codes tables in the Hebrew Bible text can also be found in an English text known to have no unusual tables other than those that occur by chance, then what is claimed to be unusual in the Hebrew Bible text is certainly not unusual and we should be skeptical of the claim that they are unusual.

But Thomas and McKay do not tell us that they have done any experiment and if they did do an experiment they do not tell us the p-value of the experiment. What we should expect is that the p-value of the experiment is uniformly distributed over the interval [0,1]. So 99% of the time we should expect to observe a p-value of 1% or higher. 1% of the time we should expect to observe a p-value of 1% or smaller. If the p-value of the experiment is greater than the significance level of the test, then we do not reject the Null hypothesis. Only if the p-value of the experiment is less than the significance level of the test do we reject the Null hypothesis and conclude that we are observing something unusual. A reasonable significance level might be 1%. For the skeptical inquirer, only observing something unusual in a setting that has nothing unusual is a cause to be skeptical.

For example, suppose the compactness of the table someone demonstrates has a p-value of 30%. This means that 30% of the time a table having ELSs of the given key words and with a compactness better than the table they demonstrated can be found in a randomly selected text from the control text population. In this case their result is entirely expected and we in fact learn nothing from their example. A test at the 1% significance level would not reject the Null hypothesis.

Let us now turn to the examples of Thomas and McKay. Thomas shows tables of word pairs such as Nazi with Hitler and Roswell with UFO. Indeed these are tables that can be found in the text of War and Peace. But do they cause us to reject the Null hypothesis of No Effect? Only if they cause us to reject the Null hypothesis of No Effect do we have cause to be skeptical.

Likewise McKay shows tables of some leaders who were assassinated with a phrase or word that is in the text of Moby Dick nearby an ELS of the person who was assassinated. The table below shows the key word pairs.

Person Assassinated	**Assassinated Key Word**
Dollfuss	Assassin
I Gandhi	The bloody deed
Moawad	An exploding bomb
ML King	To be killed
Kennedy	Had been killed
Kennedy	He shall be killed
Lincoln	Killed
Rabin	Shot
Lady Diana	Mortal in the jaws of death

The second criteria of the protocol is that the key words must be selected a priori. Now the list of people assassinated is certainly not a priori. For example, why would one select two US presidents who were assassinated instead of the four who were? Why was Garfield and McKinley omitted? Why does Gandhi and King have first initial and the others none?

There are four presidents of African Nations who were recently assassinated: Kabila, Ndadaye, Ntaryamira, and Habyarimana. Why are they omitted? What about the recent assassination of the Afghan president Qadir and the Serbian president Djindjic? Liberia had two presidents Roye and Tolbert who were assassinated. Sri Lanka's president Premadasa and Prime minister Bandaranaike were assassinated. Bangladesh's president Rahman was assassinated and Bhutto was judicially assassinated. What about Ferdinand of Austria Hungary, Ghali of Egypt, Carlos of Portugal, or Aquino of the Phillipines? To make demonstrably a priori list one cannot just subjectively select one name or another. Rather a criteria must be stated in advance and then uniformly applied to generate the list of murdered world leaders.

Also the list of assassinated key words are certainly not a priori. Many are phrases snooped in the text after examining ELSs of the murdered leader. Words obtained by snooping cannot be used. Probabilities only make sense when the entire experiment, including the key words, is specified before observing any data.

Selection of the key words in an *a priori* manner is the only guarantee that the probabilities calculated mean something. It is also the only guarantee that the experimenter did not do thousands of trials with various other key words and selected those that worked and then did their statistical calculation based on the selected key word pairs. An experiment done with selective omissions, as is the McKay replication of the famous great rabbis experiment in the Hebrew text of War and Peace, (McKay et. al. 1999) is a fraudulent experiment.

To make concrete what we are stating, we give an example of an experimental protocol.

(1) The first criteria of the protocol is the text to be tested. We select an English translation of War and Peace.

(2) The second criteria of the protocol is that the key words must be selected a priori. To make a priori list one cannot just subjectively select one name or another. Rather a criteria must be stated in advance and then uniformly applied to generate the list of murdered world leaders.

To make this a priori, we look for someone else's list of murdered world leaders. For example we could use Google to make the search using the key words: *murdered leaders*. The biggest list we find (in 2003) is at[1]

www.crikey.com.au/politics/2003/03/13-03march13murders.html.

We commit to use this list.

Consulting Roget's Thesaurus of the English Language in Dictionary Form (1938) we find the list of verbs under kill to include: killed, slain, murdered, assassinated, poisoned, butchered, slaughtered, lynched, finished, destroyed, dispatched, shot, stabbed, bayonetted. Verb phrases listed such as to shed blood or to put an end to we do not use since we are looking for verbs that do not need any additional words. Finished, destroyed, and dispatched in current English usage would probably not be employed as verbs in a newspaper story about an assassination. Nevertheless since they are listed in our source, we must commit to use them.

(3) The third part of the protocol is the determination of how the min and max skips for each key word are determined. We will set the minimum skip to be 1 and the maximum skip set so that the number of expected ELSs in a letter randomized text is just greater than 10.

(4) The fourth part of the protocol is the cylinder size search. For any pair of ELSs, cylinder sizes to be considered in the search are only those cylinder sizes that are

[1] That website in 2007 no longer exists.

resonant with one of the ELSs. Resonant means that the maximum row skip and column skip an ELS can have on a cylinder is 10.

(5) The fifth part of the protocol is the control text population. We use a control text population that is a virtual control text population. Each text of the control population has exactly all the ELSs of the given text with their starting positions chosen at random.

(6) The sixth part of the protocol is the number of texts to be sampled. We will sample 10,000 texts from the control text population. The first text we sample will be the given text. Under the Null hypothesis, it too has No Effect.

(7) The seventh part of the protocol is the definition of the compactness measure. For the purpose of this discussion the compactness measure selected is the area of the smallest area table that can be found in the sampled text using the ELS skip specification and resonant cylinder specification. The area is the number of rows times the number of columns in the table. This measure has a direct intuitive meaning but is very different from the compactness measured used by WRR. (We name the area compactness measure here only to avoid a lengthy technical discussion of compactness measures. The area measure is not among the more sensitive compactness measures, meaning that there are better ones.)

(8) The eighth part of the protocol is the specification of the hypothesis to be tested. The hypothesis we test is the Null hypothesis of No Effect in the given text. We will test this hypothesis against the alternative that there are more tables than expected by chance that are more compact than expected.

(9) The ninth part of the protocol is the significance level of the hypothesis test. We specify the significance level of the test to be .01.

(10) The tenth part of the protocol is the specification of the statistical methodology by which the p-value of the test is determined. For our specification, the statistical methodology by which the p-value of the test is determined has two levels. In the first level for each key word pair and for each text, we determine its normalized rank. The normalized rank is the count of the number of texts from the sampled texts having smaller compactness for its best table plus one half the number of texts from the sampled texts having equal compactness, the total being divided by the number of texts in the sampled text population.

The second level is the calculation of the score for each text. The score for each text is the product of the normalized ranks. The p-value of the test is then the normalized rank of the score of the given text.

It should be now clear that by merely demonstrating some code table in a text known to have no code tables other than those that occur by chance we learn nothing about the p-value and therefore cannot infer anything one way or the other. In order for the Thomas and McKay reasoning to be correct, they need to demonstrate that the claimed effect:

“small p-values for *a priori* selected logically historically related key word pairs” can be observed in the War and Peace text. If that claim were in fact true, we would be skeptical of something in the methodology or in the non a priori selection of key word pairs. Should an experiment actually be done of the type described above with the War and Peace text, we should certainly expect that the p-value would be much greater than the significance level of the test and therefore the Null hypothesis would not be rejected and we would have nothing over which to be skeptical.

This being so, we must be skeptical of Thomas’s and McKay’s reasoning that their example demonstrations of tables without experiment protocols and without p-values demonstrate that something unexpected has occurred. Rather what they have given us is itself something to be skeptical about. Their tables simply do not qualify as code tables.