Testing The Torah Code Hypothesis: The Torah Code Effect is Real

Robert M Haralick
Computer Science, Graduate Center
City University of New York
New York, New York 10016
haralick@gc.cuny.edu

Revision Date: October 14, 2003

Abstract

Witztum, Rips, and Rosenberg (1994) published a paper describing a statistical Torah Code experiment in which the equidistant letter sequences of the appellations and death/birth dates of an *a priori* set of famous rabbinic personalities formed unusually compact formations in the *Genesis* text. By a Monte Carlo experiment they showed that the probability had to be less than 16/1,000,000 that this would have happened by chance. Therefore, they concluded that this was not a chance event.

McKay et. al. (1999) argued that the Torah code experiments of Witztum, Rips, and Rosenberg on the *Genesis* text succeeded because one way or another they selectively omitted appellations to make the experiment produce a seemingly statistically significant result. To demonstrate this they cooked an experiment using a Hebrew text of *War and Peace.* They showed that by selective omissions and some spelling stretches their Monte Carlo experiment yielded a comparably small probability. In essence, they argued that had there been no stretches and had a full set of appellations been used, neither the rejection of the Null hypothesis of no Torah Code effect for the experiment using the *Genesis* text nor the rejection of the Null hypothesis for their cooked experiment using the *War and Peace* text would have happened.

In this paper we describe an experiment which proves the McKay et. al argument to be fallacious. We combine the appellation lists of McKay et. al. and Witztum et. al. to form a more complete list of appellations with no selected omissions. We designed an improved protocol for the experiment using statistically more powerful compactness measures and an ELS Random Placement control text population. We tested the Null hypothesis of no Torah code effect against four different alternative hypotheses. Our experiments show that the combined list of appellations has the same or a slightly stronger effect in the *Genesis* text than the original list for three out of the four alternative hypotheses and a slightly weaker effect for the fourth alternative hypothesis. For the McKay list in the War and Peace text there was a significant decrease in the effect for all four alternative hypotheses. And with the improved protocol, the combined list had a statistically insignificant effect in the *War and Peace* text.

These results provide evidence of the fallaciousness of McKay et. al.'s assertion that had a more full list of appellations been used, the Null hypothesis of no Torah code effect would not have been rejected in both the *Genesis* text and the *War and Peace* text. We conclude that the Torah code effect is real for the great rabbis experiment using the *Genesis* text, that there is no Torah code effect in the *War and Peace* text, and there were some valid and encoded appellations on the McKay List that were not in the original list.

Introduction

In the last few years there have been a number of books discussing Torah codes or Bible codes. Noteworthy among these have been the books of Drosnin (1997), Satinover (1999), Ingermanson (1999) and Drosnin (2003). There have been some television documentaries and a variety of different commercially available programs to find codes. Different religious organizations, Christian and Jewish, have included Torah codes in their seminars which are designed to open the doors to religious spirituality and observance to those of a secular life style. There are many websites devoted just to demonstrating either various current events or religious themes through Torah codes.

One common message of the proponents of Torah codes is how amazing and unexpected such codes are. Witztum, Rips, and Rosenberg (WRR) (1994) published the results of a scientific experiment, now commonly known as the great rabbis experiment, designed to determine in a proper statistical way whether there is a Torah code effect in the Genesis text. Their conclusion was that their results had a very small probability, no more than 16/1,000,000, of occurring by chance. That paper then stimulated a flurry of critical academic discussion about whether the WRR results were in fact real. See McKay et. al. (1999). In this paper, we analyze the Witztum et. al. (1994) experiment and the counter experiments of McKay et. al (1999) to help determine the truth of the situation.

We begin by defining equidistant letter sequences, a key concept in Torah codes. Let a text consist of a sequence of characters $c_1,\ldots,c_N$ with the spaces and punctuation marks removed. An equidistant letter sequence (ELS) of length L, skip s, and beginning position b is the character subsequence $c_b,c_{b+s},\ldots,c_{b+(L-1)s}$. The Torah code hypothesis states that historically logically related key words tend to have ELSs that are in more compact formations than expected by chance. Demonstrations of such compact formations are shown by Torah code popularizers as tables where the Torah text is spiraled around a cylinder of given number of columns and the compact ELS formations are shown in a small window from this cylinder.

McKay et. al. have argued that the great rabbis Torah code experiments of Witztum, Rips, and Rosenberg (WRR) (1994) on the *Genesis* text succeeded because one way or another they selectively omitted and added appellations to make the experiment produce a seemingly statistically significant result. Thereby WRR had to reject the Null hypothesis of no Torah code effect and had to conclude that it was not a chance

occurrence that the *Genesis* text could contain equidistant letter sequences (ELSs) of the appellations of famous rabbinic personalities paired in compact formations with ELSs of their birth or death dates. McKay et. al. demonstrate that by selective omissions and some spelling stretches and additions they could produce an experiment where they must reject the Null hypothesis of no Torah Code effect on a Hebrew text of *War and Peace*. In essence, they argue that had a full set of appellations been used without any selective omissions or additions or spelling stretches, the Null hypothesis for the experiment on the *Genesis* text would not have been rejected and the Null hypothesis of no Torah Code effect on the War and Peace text would not have been rejected.

Publicly repeatable scientific experimentation and re-experimentation is the way evidence is gathered in science. Science proceeds by doing, redoing and refining methodology and experiments. Each additional experiment is designed to provide evidence for pinning down something that might have been uncontrolled for in an earlier experiment or something which may have been less than optimal in an earlier experiment. In experiments measuring physical constants, each additional experiment is designed to reduce the variance of the measured constant. In detection and recognition experiments, each additional experiment is designed to reduce the false alarm rate and/or misdetection rate. In this sense, Torah Code experiments are no exception. What may be deficient or less than optimal in a prior experiment can be improved upon in a later refined experiment from which something more or new may be learned.

In this paper we describe a set of refined rabbi experiments using a better experimental protocol and an appellation list having no wiggle room. Our experimental results show that McKay et. al's assertion is wrong: that had a more complete and correct appellation list been used, the significance of the WRR experiment on the *Genesis* text would disappear. On the contrary, our experiments show that when a more complete appellation list is used, the statistical significance of the effect tends to stay the same or increase rather than decrease, opposite to what McKay et. al. had asserted would happen.

To set the perspective from which our new experiments were done and understand the nature of the improvement of our experimental protocols, we must first review the outstanding criticisms of the WRR experiment.

(1) the selection of appellation and dates.
(2) the compactness measure

We discuss each of these criticisms in turn.

Monkey Queries

The control population in WRR was a population of permuted pairings of appellations and dates. Each set of permuted pairings of appellations and dates sampled from the control population in a trial of the experiment is here called a monkey query. The main problem with monkey queries as WRR did it is that they permuted the pairing between sets of appellation and sets of dates instead of permuting the pairing of appellation and dates as would be done by a standard permutation protocol. Because they permuted sets

with sets, the number of appellation date pairings from trial to trial changed and then this had to be taken into account by another level of normalization through the Fisher statistic.

We use an improved protocol based on a monkey text control population. This makes every trial in the experiment completely symmetric (a feature that the WRR experiment lacked) so the p-value of the experiment has the correct statistical meaning. Furthermore, an experiment having a small p-value would directly mean that the Torah text is somehow a special unusual text in the control populations of texts.

Five years ago we began using monkey texts consisting of word permuted Torah texts. Later other varieties of permuted texts were thought about and various small experiments run. Each permuted text choice could be criticized in some way. The criticism goes along the lines that the texts in the control population could for some unknown reasons, having nothing to do with encodings or non-encodings, have ELS statistics that differ from the Torah text. A successful experiment may say nothing more than it is possible through such ELS statistics to distinguish the Torah text from the texts in the population because of such differences. But these differences may have nothing to do with the Torah Code hypothesis. The only text population that would not be subject to this kind of criticism is a text population that had exactly the same ELS statistics as the Torah text. Here ELS statistics mean the number of ELSs that each key word has and the skip intervals of the ELSs of the key words.

With respect to ELSs, the simple version of the Torah Code hypothesis states that ELSs of historically logically related key words tend to have ELSs in a more compact arrangement than would be expected by chance. Therefore, the most natural and direct way to test this hypothesis is to use a population of texts each of which has exactly the same ELSs as the Torah text, the only difference being that the ELSs are positioned randomly. Such a population we call the ELS Random Placement Population. The ELS Random Placement Population is a virtual text population and is the one we use in our new experiments. We reported about the ELS Random Placement Population at the 2000 Torah Code conference in Jerusalem. The ELS Random Placement Population is conservative in that any part of the Torah code effect due to the Torah having more than an expected number of ELSs for any key word is cancelled out. Likewise is cancelled out any chance occurrence of ELSs of different key words having resonating skips.

Appellation and Dates

The McKay et. al. criticism of the appellation and dates, particularly the appellations, is that McKay et. al. essentially claim that WRR reviewed in private many of the possible appellations for each rabbi and made sure to include in their experiment those appellations that would influence the result to be a small p-value and threw away a number of those appellations that would influence the result to be a large p-value. They argue that by peeking ahead and selecting appellations as they did, WRR made a *non-a priori* experiment. Therefore, they argue that the resulting small p-value is entirely expected and does not imply that anything unusual happened in the experiment.

To show that this is true, they did exactly this with respect to a Hebrew translation of *War and Peace*. They reviewed in private many of the possible appellations and appellation spellings for each rabbi, and made sure to include in their experiment using *War and Peace* those appellations that would influence the result to be a small p-value and threw away a number of those that would influence the result to be a large p-value. Furthermore they also threw away those appellations that contributed to a small p-value in the WRR experiment using the *Genesis* text. Thereby McKay et. al. were able to make a parlor trick in which the p-value of their cooked experiment with the *Genesis* text was large (insignificant) and the p-value of their experiment with the *War and Peace* text was small (apparently significant). See Bar-Natan and McKay (1997), Bar-Hillel et. al. (1998) and McKay et. al. (1999) for details.

What was the purpose of McKay et. al. playing these appellation games? The purpose was to show that within the WRR protocol, there was still some flexibility about which appellations to include or not include and by playing within their perceived ground of flexibility, they could make choices so as to produce a significant result in a text such as *War and Peace* which everyone would agree has only chance ELS formations. If this counterfeit result could be produced in *War and Peace*, then they argue it surely could have been produced in the *Genesis* text by the same tricks. So McKay et. al. conclude that this must have been the kind of counterfeiting that is responsible for the WRR rejection of the Null hypothesis of no Torah Code effect in *Genesis*.

The argument of McKay et. al. is fallacious in a number of ways. Firstly, just because it is possible to produce a counterfeit experiment as they did does not logically imply that the WRR experiment is counterfeit. That is, just because some money is counterfeit does not imply all money is counterfeit. Secondly, McKay et. al.'s perception of the flexibility in the WRR protocol is flawed. Witztum (2000) shows that McKay et. al. in many cases misinterpreted and/or stretched the WRR appellation and spelling protocols out of bounds in forming their appellation list.

Neither of these points however, is the most serious flaw of the McKay et. al. argument. The flaw we now discuss is so significant that once we recognize it we have no choice but to throw out their argument. Consider what has happened. McKay et. al. turned what they perceived as selections and omissions in the WRR list of appellations into a game. They argued that if WRR had the freedom to select this way or omit this way then they too could select and omit and in doing so produced the parlor trick of an apparently significant result in *War and Peace*. However, the principle behind the Torah Code experiment of Witztum, Rips and Rosenberg never contained any aspect of selective omissions of valid appellations. Indeed, if any additional valid appellations or spellings were to be discovered after the WRR experiment, WRR would argue that the proper thing to do is to redo the experiment with the new additions. That is how things would proceed in a physics experiment. Any deficiency discovered after an experiment had been done must be rectified in a new refined experiment. This is the imperative of scientific progress.

If the phenomena reported by WRR in the *Genesis* text were due to selectivity, then we expect the effect would disappear if the selective omissions were to be rectified. On the other hand, if the phenomena reported by WRR were real and there were in fact valid appellations omitted in the WRR experiment, then we would expect that a rectified appellation list would yield a stronger result in the *Genesis* text. Furthermore, because the McKay et. al. appellation list is known to have appellation omissions and appellation selections and spellings tuned to produce a significant result in *War and Peace,* we would expect that a rectified list should produce a less significant result in the *War and Peace* text.

Therefore, let us not argue, as Witztum has done, about the details of the additional appellations or appellation spellings that McKay et. al. used. We take everything in the McKay list, even those that Witztum has argued are incorrect appellations or incorrect spellings. Let us just take what McKay did as an indication that in fact not all the reasonable appellations for the rabbis were used in the WRR experiment. So we will rectify that. We will use the research of McKay et. al. to remove any appellation omissions and appellation selections that might have been present in the WRR experiment. We will do this by merging the two appellation lists together. This merging does not completely rectify the combined list of either incorrect appellations or appellation spelling stretches tuned to the *War and Peace* text.. But to make the experiment one without wiggle room, we accept for now the combined list.

If in a refined experiment on the *Genesis* text using the combined lists, the Null hypothesis of no Torah Code effect cannot be rejected, then McKay et. al.'s own work of making public the additional appellations would be turned on its head and provide evidence supporting the Torah Code hypothesis. Rejection of the Null hypothesis would show the fallaciousness of their argument that had a more complete list of appellations been used, the effect would disappear in the *Genesis* text. It is just this rejection of the Null hypothesis that we will demonstrate by our experiments.

The table below shows the resulting merged appellation set.

Index	***Rabbi's Identity***	***Combined WRR and McKay Appellation List***
1*	Abraham of Narbonne	רביאברהם, הראבי, הרבאבד, הראבד, האשכול
2*	Abraham Yitzchaki	רביאברהם, יצחקי, זעאברהם
3	Abraham HaMalakh	רביאברהם, המלאך
4*	Aharon of Karlin	רביאהרן
5	Eliezer Ashkenzai	מעשיהשם, מעשייהוה, מעשיה, בעלמעשיה
6*	David Oppenheim	רבידוד, אופנהים
7*	David Nieto	רבידוד, דודניטו
8	Chaim Abulafia	רביחיים, מהרחא, המהרחא
9	Chaim Benbenest	בנבנשת, בנבנשתי, הרבחביב, הרבהחביב, רבחביב, רביחיים
10	Chaim Capusi	רביחיים, כפוסי, כאפוסי, בעלנס, בעלהנס

11*	Chaim Shabetai	רביחיים, חיימשבתי, מהרחש, המהרחש
12*	Yair Chaim	חותיאיר
13	Yehuda Chasid	רבייהודה, הריחסיד, יהודהסגל
14	Yehuda Ayash	רבייהודה, מהריעיאש, עאיאש
15*	Yehoseph HaNagid	רבייהוסף
16*	Yehoshua of Cracow	רבייהושע, מגנישלמה
17	The Maharit	רבייוסף, מטרני, מטראני, טראני, יוספטרני, ריטראני מהרימט, המהרימט, המהריט, הריטרני, מהריט, ריטרני, הריטראני
18	Yoseph Teomin	רבייוסף, תאומים, פריצגדצים
19*	Yaakov Beirav	רבייעקב, יעקבבירב, מהריבירב, הריבר
20	Israel Yaakov Hagiz	חאגיז, בעלהלקט, ריחגיז, מהריחגיז
21*	The Marahal	רבייעקב, מולין, יעקבסגל, יעקבהלוי, מהריסגל, מהריהלוי, מהריל, המהריל
22	The Yaabetz	הריעבץ, עמדין, הריעמדן, הריעמדין, ריעמדין, היעבץ
23	Yitzchak Horowitz	רבייצחק, הורוויץ, הורוביץ, יצחקהלוי
24	Menachem Krochmal	רבימנחם, קרוכמל, קרוכמאל, רבימענדל, צמחצדק
25	Moshe Zacuto	רבימשה, זכותא, זכותו, משהזכות, משהזכותו, המזלן משהזכותא, מהרמזכות, מהרמז, המהרמז, קולהרמז
26	Moshe Margalit	רבימשה, מרגלית, פנימשה, מרגליות
27*	Azariah Figo	רביעזריה
28	Immanuel Hai Ricchi	אוהבורע, הונעשיר, העשיר, אחהער, ישרלבב
29	Shalom Sharabi	רבישלום, מזרחי, שרעבי, שרשלום, מהרשש, המהרשש
30	Shlomo of Chelm	רבישלמה, חעלמא, שלמהחלמא
31	Meir Eisenstat	רבימאיר, איזנשטאט, איזנשטט, מהרמאש

Appellations for the combined WRR McKay list. Indexes with * indicate that these appellations are identical for the WRR list and the McKay list.

There are 13 rabbis for which the appellations on the McKay list and the WRR list are identical. WRR includes rabbi 18 while McKay excludes rabbi 18. McKay includes rabbi 31 while WRR excludes rabbi 31. The first table below shows the 23 appellations included in the WRR list that are excluded in the McKay list. The second table below shows the 31 appellations included in the McKay list that are excluded in the WRR list. So of the 97 appellations of WRR, McKay removed 23 and added 31, a change of 54 appellations out of the 97 appellations. In no way is this a "small change" as stated by McKay on his website http://cs.anu.edu.au/~bdm/dilugim/torah.html.

Index	*Rabbi's Identity*	*In WRR List But Not McKay List*
3	Abraham HaMalakh	המלאך
9	Chaim Benbenest	בנבנשת
10	Chaim Capusi	כפוסי, בעלנס, בעלהנס
17	The Maharit	המהרימט
18	Yoseph Teomin	רבייוסף, תאומים, פריצגדצים
20	Israel Yaakov Hagiz	חאגיז
22	The Yaabetz	הריעמדן
23	Yitzchak Horowitz	הורוויץ
24	Menachem Krochmal	קרוכמל
25	Moshe Zacuto	משהזכותא, זכותא, זכותו, משהזכותו
28	Immanuel Hai Ricchi	אחהער, ישרלבב
29	Shalom Sharabi	המהרשש, מזרחי, שרשלום, מהרשש

Appellations included in the WRR list and excluded in the McKay list.

Index	*Rabbi's Identity*	*In McKay List But Not WRR List*
5	Eliezer Ashkenzai	מעשיה, בעלמעשיה
8	Chaim Abulafia	המהרחא, מהרחא
9	Chaim Benbenest	רבחביב, בנבנשתי, הרבחביב, הרבהחביב
10	Chaim Capusi	כאפוסי
13	Yehuda Chasid	הריחסיד, יהודהסגל
14	Yehuda Ayash	עאיאש
17	The Maharit	ריטראני, ריטרני, הריטרני, הריטראני
20	Israel Yaakov Chagiz	ריחגיז, מהריחגיז
22	The Yaabetz	ריעמדין
23	Yitzchak Horowitz	הורוביץ
24	Menachem Krochmal	קרוכמאל
26	Moshe Margalit	מרגליות
28	Immanuel Hai Ricchi	אוהבורע, הונעשיר, העשיר
30	Shlomo of Chelm	חעלמא, שלמהחלמא
31	Meir Eisenstat	רבימאיר, איזנשטאט, איזנשטט, מהרמאש

Appellations included in the McKay list and excluded in the WRR list.

The Compactness Measure

Four years ago we seriously began exploring different choices of compactness measures. This exploration was initially driven by a sense of aethestics more than anything else. Without going into all the details (see Witztum, Rips, and Rosenberg, 1994), the heart of the WRR compactness measure between a pair of key words is a sum they called *Omega.* To normalize this *Omega* value, they do some ELS position perturbations and convert the *Omega* value to what they call a *c*-value. The p-value of the WRR experiment is obtained by combining the *c*-values over all appellation date pairs to one grand value. The relative rank of this grand value for the correct pairing of the appellation and dates was the p-value of the experiment. Our criticism of the compactness measure was twofold: (1) that the *c*-value normalization was cumbersome and unaesthetic; and (2) it did not simply measure the best most compact ELS arrangements.

The simplest compactness measure is a 1D measure: the length of the smallest text segment having at least one ELS of each of the key words of a key word set. If this measure would be the most powerful compactness measure, the need to choose cylinder sizes as required by the WRR protocol would go away. Unfortunately our early statistical measurements only using the WRR rabbis list one showed that this compactness measure was not as good as the *Omega* compactness measure of WRR. We explored other compactness measures, some of which were explored by McKay et. al. and some of which were not. Of those that we explored, among the compactness measures that had good false alarm and misdetection statistics was the class of measures we call here as H_1, H_2, H_3, and H_4. They are defined as follows.

Let an appellation date key word pair be (w_1,w_2) and let a text index be t. For each key word w, let $E(w,t)$ be the set of ELSs of w in text t. For any pair (e_1,e_2) of ELSs let $C(e_1,e_2)$ be the set of cylinder sizes that resonate with e_1 or e_2. Here we define a cylinder size to resonate with an ELS if the row skip of the ELS on the cylinder is 10 or less and column skip of the ELS on the cylinder is differs by no more than one column from the natural skip of the ELS on the cylinder. Let $d_{min}(e_1,e_2,c,t)$ be the closest squared distance on the cylinder of size c between the letters of ELSs e_1 and e_2 in text t. Let $d_{max}(e_1,e_2,c,t)$ be the furthest squared distance on the cylinder of size c between the letters of ELSs e_1 and e_2 in text t. For any ELS e and cylinder size c and text t let $s(e,c,t)$ be the sum of the squared row skip and squared column skip of ELS e on a cylinder of size c in text t. Then the compactness measure H_1 for key word pair (w_1,w_2) on text t is a function of the terms

$$\{\, d_{min}(e_1,e_2,c,t)[\, s(e_1,c,t) + s(e_2,c,t)\,] \mid e_1 \text{ in } E(w_1,t),\ e_2 \text{ in } E(w_2,t),\ c \text{ in } C(e_1,e_2)\,\}$$

There are a number of functions that could be chosen, each with different statistical efficiency and power. For example we could take the geometric mean, arithmetic mean or the harmonic mean. In a talk given at the Torah Code conference in 2000, we showed that the harmonic mean, which was essentially what WRR used to combine ELS pairs of a given appellation date pair, was statistically better.

The harmonic mean h of N positive numbers $x_1,...x_N$ is defined by

$$h(x_1,...x_N) = N/(1/x_1 + ... + 1/x_N)$$

Let us call the resulting measure for appellation date pair (w_1,w_2) and text t $H_1(w_1,w_2,t)$. For each monkey text t of the experiment and for each appellation date pair (w_1,w_2) of the key word pair set Wp of the given rabbi, we observe $H_1(w_1,w_2,t)$. For each (w_1,w_2) pair, the raw data of the experiment is the list of values $H_1(w_1,w_2,0),\ldots, H_1(w_1,w_2,T)$. For each pair (w_1,w_2) these values are rank normalized over all t. Rank normalization here means that each value v of the list is replaced by a fraction whose numerator is the number of values in the list less than v plus one half the number of values in the list equal to v and whose denominator is the number of values in the list. Then, following how WRR combined the harmonic means across all appellation date pairs, we compute $G_1(Wp,t)$ for the text t and a word pair set Wp as the geometric mean of the rank normalized values of $H_1(w_1,w_2,t)$ taken over all word pairs (w_1,w_2) in the set Wp. The p-value, $P_1(Wp,t)$ for a text t is then just the rank normalized value of $G_1(Wp,t)$ taken over all texts t. As the *Genesis* text is the first text $t=0$, the p-value associated with the compactness measure H_1 is the rank normalized value $P_1(0,Wp)$. H_2 is defined similarly to H_1 using the terms from

$$\{\ d_{max}\,(e_1,e_2,c,t)[\ s(e_1,c,t) + s(e_2,c,t)\ ]\ |\ e_1 \text{ in } E(w_1,t),\ e_2 \text{ in } E(w_2,t),\ c \text{ in } C(e_1,e_2)\ \}$$

If instead of including in the harmonic mean all the terms from the resonanting cylinders, we only include the best cylinder, then we have the compactness measure H_3. It is defined using terms from

t

As before, we select the function to combine all these terms to be the harmonic mean. As for $H_1(Wp,t)$, $H_3(w_1,w_2,t)$ is rank normalized over all monkey texts. The geometric mean of each rank normalized value is taken over all appellation date pairs to form $G_3(t,Wp)$. The p-value $P_3(0,Wp)$ associated with the compactness measure H_3 is the respective rank normalized value of $G_3(0,Wp)$.

The compactness measure H_4 is defined in a similar way to H_3 except using the terms from

$$\{\ \min_{c \text{ in } C(e_1,e_2)}\ d_{max}\,(e_1,e_2,c,t)[\ s(e_1,c,t) + s(e_2,c,t)\ ]\ |\ e_1 \text{ in } E(w_1,t),\ |\ e_2 \text{ in } E(w_2,t)\ \}$$

The initial experiments using H_1, H_2, H_3, and H_4 along with other less well performing measures were reported at the 2000 Torah Codes Conference in Jerusalem. At that time we demonstrated that this class of measures were better measures than the WRR *Omega* measure and measures of best area, perimeter, maximal side length, or diagonal of best table on best cylinder.

It is interesting to compare our measures with that of WRR. The kernel of the terms we use in our measures bear some similarity to those used by WRR. WRR's kernel is of the form $d_{min}\ (e_1,e_2,c,t) + s(e_1,c,t) + s(e_2,c,t)$. The problem with this kernel is that if the optimal scale factors that should go with $d_{min}\ (e_1,e_2,c,t)$ are different from the optimal scale factors that should go with $s(e_1,c,t) + s(e_2,c,t)$, then the choice of simply adding will make a less than optimum measure. By multiplying the distance term with the skip terms we do not have to contend with the problem of disparate scales since the multiplied term simply retains the product of the scale factors.

How The Four Compactness Measures Were Chosen

Our exploratory studies included 32 different compactness measures. We measured the p-value of each of these 32 measures against each of the 32 rabbis of WRR list one. The measures H_2 and H_4 were among the top three measures using the geometric mean for combining the p-values and were the top two measures using the best star team approach for combining the p-values. Both measures had an overall p-value an order of magnitude or more smaller than the WRR *Omega* measure. Before deciding to use measures H_2 and H_4 in our new experiments we examined the compactness measure p-values rabbi by rabbi and noticed that when these compactness measures tended to be small, it was not unusual for the measures H_1 and H_3 tended to be large and when these measures tended to be large it was not unusual for the H_1 and H_3 compactness measures to be small. We then checked every pair of correlations between the compactness values and saw that H_1 had minimum correlation with H_4 and H_2 had minimum correlation with H_3 and that H_3 and H_4 had minimum correlation with each other. All this suggested that H_2 and H_4 sometimes worked in a symmetric push pull fashion with H_1 and H_3: where one was good the other was bad and visa-versa. Finally, we set a running threshold p from .01 to .26 on the p-value of each rabbi for each compactness measure and for each pair of the 32 compactness measures counted the number of rabbis from list one that had compactness less than or equal to p by one of the measures of the pair. We observed that from thresholds p=.1 to p=.14 the measure pair H_3 and H_4 had a total of 16 rabbis having a p-value less than or equal to p on one of the measures H_3 and H_4 and no other measure pair performed better. This reinforced our observation that although H_2 and H_4 worked better overall, there were some rabbis whose appellation and date encodings were captured better by compactness measures H_1 and H_3. Thus from our exploratory study on WRR rabbis list one, we decided to use compactness measures H_1, H_2, H_3, and H_4 for our experiments involving the WRR and McKay appellation and dates for rabbis list two.

The Experimental Protocol

A Torah Code experiment has a protocol consisting of the ELS search protocol, a cylinder size protocol, a control population protocol, a compactness measure protocol, and an hypothesis testing protocol. In our experiment, the ELS search protocol consists of having the minimum ELS skip be one and maximum ELS skip set for each key word so that the expected number of ELSs in a text chosen at random from a letter permuted population would be ten.

The cylinder size protocol considers for each pair of ELSs only those cylinder sizes such that one ELS of an ELS appellation date pair would have a cylinder row skip of less than or equal to 10 and a cylinder column skip of less than or equal to 1. The 10 here functions in the way WRR divides an ELS skip by the integers 1 through 10 to obtain the cylinder sizes. However, our criteria of a cylinder column skip of less than or equal to 1 makes the ELS skip resonate stronger with the cylinder size than in the WRR protocol.

To test the Torah Code hypothesis, we posit the Null hypothesis of no Torah Code effect. Under this hypothesis, the positioning of the ELSs of the key words in a text is uniformly distributed. After all what else can it mean that there is no compact grouping of ELS arrangements of corresponding appellations and dates? So the control population of texts we use is the ELS Random Placement population. The compactness measure protocol is P_1, P_2, P_3, and P_4.

An hypothesis test of the Null hypothesis must be done against some alternative hypothesis. For us, the alternative hypothesis is that the there is a Torah Code effect. But depending on exactly what we mean by "there is a Torah Code effect" our hypothesis test will be different. There are four alternative hypotheses we are interested in:

(1) the appellation date compactness values tend to be smaller in the *Genesis* text than in the texts of the control population;
(2) there are more appellation date compactness values that are small in the *Genesis* text than in the texts of the control population;
(3) the rabbi compactness values tend to take smaller values in the *Genesis* text than in texts of the control population;
(4) there are more rabbi compactness values that are small in the Genesis text than in the texts of the control population.

We test alternative (1) by defining Wp to be the set of all appellation date pairs taken over all rabbis. Our observed data has one record per text t where each record consists of the tuple

$$< (H_1(w_1,w_2,t), H_2(w_1,w_2,t), H_3(w_1,w_2,t), H_4(w_1,w_2,t)) \quad | (w_1,w_2) \text{ in } Wp >$$

We independently rank normalize each of the fields in the data to form records

$$< (R_1(w_1,w_2,t), R_2(w_1,w_2,t), R_3(w_1,w_2,t), R_4(w_1,w_2,t)) \quad | (w_1,w_2) \text{ in } Wp >$$

After taking geometric means over all (w_1,w_2) in Wp a record for text t consists of the four-tuple

$$(G_1(Wp,t),\ G_2(Wp,t),\ G_3(Wp,t),\ G_4(Wp,t))$$

Each of the four fields is independently rank normalized to produce the rank normalized records. For text t, the rank normalized record consists of the four-tuple

$$(P_1(Wp,t),\ P_2(Wp,t),\ P_3(Wp,t),\ P_4(Wp,t))$$

The minimum normalized rank for each record is then computed

$$P_{min}(Wp,t)=min\{ P_1(Wp,t),\ P_2(Wp,t),\ P_3(Wp,t),\ P_4(Wp,t)\}$$

The p-value p of the experiment is the ratio of the number of texts whose $P_{min}(Wp,t)$ is less than or equal to $P_{min}(Wp,0)$ divided by the total number of texts.

Note that by forming the minimum $P_{min}(Wp,t)$, trial by trial, we improve upon estimating an upper bound on the p-value by the use of the Bonferonni inequality, the methodology employed by WRR. To understand this difference, first let us understand what the p-value of the experiment is suppose to mean. The p-value means the probability that a randomly sampled text from the text population would have one of the four measures yield a rank normalized compactness value as small or smaller than observed in the first text of the experiment. Now consider how WRR would have approached our situation. There are four compactness measures so there are four experiments yielding respective p-values $P_1(Wp,0),\ P_2(Wp,0),\ P_3(Wp,0),\ P_4(Wp,0)$. As per Bonferonni an upper bound on the p-value p is

$$4\ min\{ P_1(Wp,0),\ P_2(Wp,0),\ P_3(Wp,0),\ P_4(Wp,0)\}$$

This is the value they would have calculated. The Bonferroni inequality states that

$$p \leq 4\ min\{ P_1(Wp,0),\ P_2(Wp,0),\ P_3(Wp,0),\ P_4(Wp,0)\}$$

with equality if and only if $T_1,T_2,T_3,$ and T_4 are mutually exclusive, where

$$T_n= \{\ t \mid\ P_n(Wp,t)\ <=\ P_{min}(Wp,0)\ \},\ \ n=1,2,3,4$$

When $T_1,T_2,T_3,$ and T_4 are not mutually exclusive, the case that is surely true if the compactness measures are correlated, then the true p-value is strictly less than the Bonferroni bound.

Under the symmetry conditions under which each experiment is conducted and under the assumption that the Null hypothesis is true, the observed p-value of the experiment is uniformly distributed. Therefore, a significance test of .001 means that if we were to repeat the experiment many many times for a fraction 1/1000 of the time, we will test the Null hypothesis under conditions that the Null hypothesis is true and we will mistakenly reject the Null hypothesis. The significance level of the test is our false alarm error rate given that the Null hypothesis is true.

If the p-value of the experiment is less than the significance level .001 we reject the Null hypothesis of no Torah Code effect against the alternative hypothesis that the appellation date compactness values tend to be smaller in the *Genesis* text than in the texts of the control population.

For testing the Null hypothesis against alternative hypothesis (2), for each text t we will simply count the number of word pairs (w_1,w_2) for which one of the rank normalized values of $H_1(w_1,w_2,t)$, $H_2(w_1,w_2,t)$, $H_3(w_1,w_2,t)$, $H_4(w_1,w_2,t)$ is less than .05. The p-value of the test will be the number of texts whose count is less than or equal to the count for the Genesis text divided by the total number of texts. If the p-value of the experiment is less than the significance level .001 we reject the Null hypothesis of no Torah Code effect against the alternative hypothesis that there were more smaller valued appellation date compactness values in the *Genesis* text than in the texts of the control population.

For testing the Null hypothesis against alternative hypothesis (3), we proceed exactly like alternative hypothesis (1) except instead of *Wp* being the set of all appellation date pairs taken over all rabbis, each rabbi has its own *Wp*. In this way for each rabbi, we may compute one p-value associated with the set of appellation date pairs for the rabbi. Because there is no appellation date pair common to more than one rabbi, (although there are some common appellations and some common dates) and because the rabbi p-value depends on a few appellation date pairs, we may safely assume that these p-values are approximately independent. Under the Null hypothesis, the p-values are uniformly distributed over [0, 1/T, 2/T,…,1], where T is the number of texts. For large T this is approximately uniformly distributed in the interval [0,1]. We compute the product x_0 of these p-values. Under the Null hypothesis and the independence of p-value assumption, the probability that such a product X of the p-values would be less than the observed product x_0 is given by

$$Prob\ (X < x_0) = x_0[1 - \ln x_0 + (-\ln x_0)^2/2! + \ldots + (-\ln x_0)^{N-1}/(N-1)!\ ]$$

If the p-value of the experiment is less than the significance level .001 we reject the Null hypothesis of no Torah Code effect against the alternative hypothesis that the there are more appellation date compactness values that are small in the *Genesis* text than in the texts of the control population.

For testing alternative hypothesis (4) we will count the number K_0 of p-values less than or equal to .25 Under the Null hypothesis and the independence of p-value assumption, the

number K of p-values less than or equal to .25 is distributed as a Binomial variate $B(.25,N)$, where N is the number of rabbis. If the probability that K is greater than or equal to K_0 is less than the significance level .001, then we will reject the Null hypothesis in favor of the alternative hypothesis that there are more rabbi compactness values that are small in the *Genesis* text than in the texts of the control population.

The Experimental Results

In this section we test the Null hypothesis against each of the four alternative hypotheses in accordance with the protocol for hypothesis testing discussed in the previous section. Our results show that for the combined list, at the significance level of .001, the Null hypothesis of no Torah code effect for the *Genesis* text must be rejected against all alternative hypotheses. We therefore conclude, consistent with Witztum et. al. (1994), that in the *Genesis* text it was unlikely by chance that

(1) the observed compactnesses of the appellation date pairs among all the rabbis would be as small as they were observed to be;
(2) the number of appellation date pairs having small values among all the rabbis would be as large as they were observed to be;
(3) the observed p-values of the rabbis would be as small as they were observed to be;
(4) the number of rabbis having a small p-value would be as large as they were observed to be.

Furthermore, our results show that we cannot reject the Null hypothesis of no Torah code effect at a significance level of .001 for the combined list in the *War and Peace* text against alternatives (3) and (4).

Test of Null Hypothesis Against Alternative (1)

The p-value for the test of the Null hypothesis against alternative hypothesis (1) that the appellation date compactness values tend to be smaller in the *Genesis* text than in the texts of the control population are given in the table below. For information purposes we show the p-value associated with each different compactness measure. The p-value of the hypothesis test as specified in our hypothesis testing protocol is given in the last column.

For the *Genesis* text, the hypothesis of no Torah Code effect is rejected at the significance level of .001 for all lists. For the *War and Peace* text, the hypothesis of no Torah Code effect is rejected at the significance level of .001 for the McKay list and the Combined List. We note that the p-value for the combined list in the *War and Peace* text is larger (less significant) than the McKay list indicating a weakening of the effect with the combined list.

Recall that when the McKay List was run using the WRR protocol on the *Genesis* text the Null hypothesis could not be rejected. This was because McKay et. al. had selectively

List	***Text***	***H_1 P-value***	***H_2 P-value***	***H_3 P-value***	***H_4 P-value***	***Experiment P-value***
WRR List One	*Genesis*	813/100,000	6/100,000	832/100,000	25/100,000	18/100,000
WRR List Two	*Genesis*	1/100,000	718/100,000	1/100,000	968/100,000	2/100,000
McKay List	*Genesis*	3/100,000	12059/100,000	2/100,000	9303/100,000	6/100,000
Combined List	*Genesis*	1/100,000	2402/100,000	1/100,000	2934/100,000	2/100,000
WRR List One	*War and Peace*	3677/10,000	2508/10,000	2623/10,000	4376/10,000	4384/10,000
WRR List Two	*War and Peace*	700/10,000	2214/10,000	411/10,000	2022/10,000	972/10,000
McKay List	*War and Peace*	1/100,000	17/100,000	1/100,000	119/100,000	3/100,000
Combined List	*War and Peace*	8/100,000	103/100,000	4/100,000	442/100,000	12/100,000

Table of p-values for different lists and texts and compactness measures testing the Null hypothesis against the alternative that the compactness for the given text tended to have more smaller values than expected by chance. Notice that the Null hypothesis of no Torah code effect must be rejected for the McKay list for both the Genesis text and the War and Peace text and for the WRR and combined lists for the Genesis text. Also notice that p-value of the combined list for the War and Peace text is weaker than the p-value of the McKay list for the War and Peace text.

omitted appellations that made the effect go away for the *Genesis* text under the WRR protocol. But since the effect remains under our improved protocol, it must be that the McKay list has some valid appellations that are encoded in the *Genesis* text. This result is consistent with Witztum's analysis *New Statistical Evidence for a Genuine Code in Genesis* (2000). Next notice that the p-value for the combined list in *Genesis* is smaller than the McKay list indicating that the effect is stronger in the combined list than in the McKay list. This is because the combined list has more encoded appellation date pairs than the McKay list. But because the McKay list contains some bogus appellations or appellation spellings, diluting the effect, the combined list performs similarly to as the WRR list on the *Genesis* text: a p-value of 2/100,000. Finally we notice that H_1 and H_3, measures that key in on the closest letter to letter distance between ELSs, have a better overall performance for Rabbis list two and the H_2 and H_4, measures that key in on the furthest letter to letter distance between ELSs, have a better overall performance for Rabbis list one. We shall see when we examine the test of the Null hypothesis against alternative (3) why this is so.

Test of Null Hypothesis Against Alternative (2)

The p-values for the test of the Null hypothesis against alternative hypothesis (2) that there tends to be a larger number of small compactness appellation date pairs in the *Genesis* text than in the texts of the control population are given in the table below.

In the WRR Rabbis list 2, there were 20 appellation date pairs whose rank normalized value of $min\{ H_1(w_1,w_2,0), H_2(w_1,w_2,0), H_3(w_1,w_2,0),H_4(w_1,w_2,0)\}$ is less than the p-value threshold of .05. The expected number is 8.15. For the McKay list, there were 17 such appellation date pairs. The expected number is 8.95. And for the combined list there were 24. The expected number is 11.1. All three lists had a higher number of appellation pairs than expected. All but the McKay list was statistically significant at the .05 level on the *Genesis* text. None were significant for the *War and Peace* text.

List	***Text***	***# Successful appellation date pairs***	***Total # appellation date pairs***	***# appellation date pairs both having ELSs***	***P-value***
WRR List One	*Genesis*	20	278	158	1.9×10^{-4}
WRR List Two	*Genesis*	20	298	163	2.5×10^{-4}
McKay List	*Genesis*	17	337	179	8.7×10^{-3}
Combined List	*Genesis*	24	403	222	4.2×10^{-4}
WRR List One	*War & Peace*	9	278	190	6.14×10^{-1}
WRR List Two	*War & Peace*	13	298	178	1.09×10^{-1}
McKay List	*War & Peace*	22	337	197	3.60×10^{-4}
Combined List	*War & Peace*	23	403	245	3.20×10^{-3}

Table of successful appellation date pairs. An appellation date pair (w_1,w_2) is successful if the rank normalized value of $min\{ H_1(w_1,w_2,0), H_2(w_1,w_2,0), H_3(w_1,w_2,0),H_4(w_1,w_2,0)\}$ taken over all the sampled texts is less than or equal to .05. Notice that the Null hypothesis of No Torah code effect for the Genesis text must be rejected for the original and combined lists and cannot be rejected for the McKay list.

These results are robust under a change of p-value threshold. For the combined list in the *Genesis* text, for all p-value thresholds between .03 and .24 the number of appellation date pairs whose rank normalized value of $min\{ H_1(w_1,w_2,0), H_2(w_1,w_2,0), H_3(w_1,w_2,0),H_4(w_1,w_2,0)\}$ is less than the p-value threshold is statistically significant at a significance level of less than .001. For the WRR list two in the Genesis text, for all p-value thresholds between .035 and .3, the number of appellation date pairs whose rank normalized value of $min\{ H_1(w_1,w_2,0), H_2(w_1,w_2,0), H_3(w_1,w_2,0),H_4(w_1,w_2,0)\}$ is less than the p-value threshold is statistically significant at a significance level of less than .001.

We also observe that the p-value that would be calculated assuming that the number of successful appellation date pairs is Binomially distributed B(.05,N), where N is the number of appellation date pairs both having ELSs, is not far off from the experimentally observed p-value. This indicates that the success of one appellation date pair is nearly independent of the success of another appellation date pair. For example, the probability that a Binomially distributed variate B(.05,222) would take the value of 24 or more as observed for the combined list in the *Genesis* text is 3.55×10^{-4}. The experimentally observed p-value, which does not make any independence assumption, is 4.2×10^{-4}.

Test of Null Hypothesis Against Alternative (3)

In accordance with the above testing protocol, the experiment was run rabbi by rabbi on 1,000 ELS random placement texts. The result of each experiment is a p-value. The table of p-values for the combined Rabbis list two is given below. For informational purposes we show in the table the p-values for each of the four individual compactness measures.

Rabbi Index	***Rabbi's Identity***	***H_1 P-value***	***H_2 P-value***	***H_3 P-value***	***H_4 P-value***	***Exp. P-value***
1	Abraham of Narbonne	.0915	.0035	.0015	.0015	.005
2	Abraham Yitzchaki	.3075	.9505	.4525	.9675	.501
3	Abraham HaMalakh	.2655	.2855	.1635	.4065	.266
4	Aharon of Karlin	.5000	.5000	.5000	.5000	1.00
5	Eliezer Ashkenzai	.9425	.2415	.9565	.0985	.207
6	David Oppenheim	.0415	.0975	.0335	.0475	.074
7	David Nieto	.6815	.0525	.7995	.0655	.114
8	Chaim Abulafia	.0005	.1925	.0015	.2385	.004
9	Chaim Benbenest	.4535	.6785	.4815	.7195	.690
10	Chaim Capusi	.0005	.7165	.0055	.8735	.003
11	Chaim Shabetai	.0615	.9405	.0355	.8575	.067
12	Yair Chaim	.5595	.9115	.6685	.9265	.770
13	Yehuda Chasid	.0805	.0815	.0805	.2655	.159
14	Yehuda Ayash	.6155	.9425	.7415	.9445	.808
15	Yehoseph HaNagid	.5000	.5000	.5000	.5000	1.00
16	Yehoshua of Cracow	.5000	.5000	.5000	.5000	1.00
17	The Maharit	.6985	.8195	.4485	.8565	.665
18	Yoseph Teomin	.1395	.8555	.3855	.9075	.248
19	Yaakov Beirav	.0765	.0725	.1315	.0515	.120
20	Israel Yaakov Chagiz	.1415	.3315	.2795	.4795	.275
21	The Marahal	.1565	.0435	.1995	.1095	.103
22	The Yaabetz	.2445	.7355	.0835	.6095	.182
23	Yitzchak Horowitz	.9975	.9955	.9965	.9945	1.00
24	Menachem Krochmal	.0095	.1585	.0395	.1225	.025
25	Moshe Zacuto	.0095	.6765	.0105	.6765	.025
26	Moshe Margalit	.6745	.2015	.5585	.2969	.378
27	Azariah Figo	.5000	.5000	.5000	.5000	1.00
28	Immanuel Hai Ricchi	.1005	.0215	.1365	.0425	.052
29	Shalom Sharabi	.0635	.0895	.0545	.0845	.122
30	Shlomo of Chelm	.1835	.9295	.1415	.6775	.290
31	Meir Eisenstat	.1025	.3455	.1385	.3955	.214

P-values for the 31 rabbis of the combined WRR and McKay appellation lists on the *Genesis* text for each compactness measure.

These are there just to help us understand which compactness measures are carrying the information and whether the encoding of different rabbis might be carried by different compactness measures. Consistent with our observation on WRR rabbis list one, we also observe here a push pull effect of compactness measures H_1 and H_3 versus measures H_2 and H_4. In accordance with our hypothesis testing protocol, the p-value for the experiment for each rabbi is shown in the last column.

Here we are interested in how unexpected it is for the observed 31 p-values to be as small as they were observed to be in the combined list. We test the Null hypothesis against the alternative that the observed rabbi by rabbi p-values are smaller than expected by chance.

For informational purposes, the table below shows the p-value of the hypothesis test of the Null hypothesis against the alternative that the observed rabbi by rabbi p-values are smaller than expected by chance. The p-values are shown for each compactness measure and in the last column for the measure we test against in accordance with our hypothesis testing protocol.

For the combined list, the probability that the product of the experimentally observed p-values would be as small or smaller than they were observed to be is 1.17×10^{-4}. Therefore, at the .001 significance level, we reject the Null hypothesis against the alternative hypothesis that for the combined list, there are more rabbi compactness values that are small in the *Genesis* text than in the texts of the control population.

Appellation List	***Text***	H_1	H_2	H_3	H_4	***P-value***
WRR List One	*Genesis*	9.73×10^{-3}	2.30×10^{-4}	7.27×10^{-3}	3.63×10^{-4}	4.66×10^{-3}
WRR List Two	*Genesis*	1.02×10^{-4}	1.54×10^{-2}	8.47×10^{-5}	1.18×10^{-2}	1.77×10^{-4}
McKay List	*Genesis*	1.43×10^{-3}	2.93×10^{-1}	2.92×10^{-4}	2.56×10^{-1}	1.17×10^{-2}
Combined List	*Genesis*	2.92×10^{-6}	4.68×10^{-2}	7.89×10^{-6}	6.33×10^{-2}	1.17×10^{-4}
WRR List One	*War and Peace*	4.13×10^{-1}	4.47×10^{-1}	3.19×10^{-1}	5.31×10^{-1}	3.89×10^{-1}
WRR List Two	*War and Peace*	5.25×10^{-2}	4.03×10^{-1}	3.87×10^{-2}	4.71×10^{-1}	2.01×10^{-1}
McKay List	*War and Peace*	5.13×10^{-5}	5.59×10^{-4}	1.13×10^{-4}	7.38×10^{-3}	4.88×10^{-4}
Combined List	*War and Peace*	3.67×10^{-3}	1.23×10^{-2}	3.73×10^{-3}	4.95×10^{-2}	4.25×10^{-3}

Table of $Prob\ (X < x_0)$ where x_0 is the product of the observed p-values taken over all the rabbis and X is such a product under the Null hypothesis. Notice that the Null hypothesis of no Torah code effect in the Genesis text for original and combined lists must be rejected. Also notice that p-value of the combined list for the War and Peace text is weaker than the p-value of the McKay list for the War and Peace text.

Here also, the p-value for the combined list with the *Genesis* text is smaller than the p-values for the WRR list and for the McKay list, indicating again that the combined list has stronger encodings than either the WRR list or the McKay list. The McKay list, which was selectively tuned to succeed on the *War and Peace* text, succeeds here too. However, despite some of the stretched spellings and incorrect appellations of the McKay list that is included in the combined list, notice that the Null hypothesis is not rejected for

the combined list for the *War and Peace* text, indicating that there is no Torah code effect in the *War and Peace* text.

Test of Null Hypothesis Against Alternative (4)

Here we test the Null hypothesis against the alternative hypothesis that the number of rabbis having significantly more compact arrangements of ELSs of their appellation and date pairs is larger than expected by chance. To determine whether any particular rabbi is encoded we must set a significance level of the test for that rabbi. The significance level of this test cannot reasonably be .001 since our expectation is that if a rabbi is encoded, only one or perhaps a few of the many appellation date pairs associated with the rabbi are encoded. The unencoded appellation date pairs will tend to make the result less significant. Since we choose .05 as the significance level for the test associated with any particular appellation date pair, we set a higher significance level of .25, in order that we may detect the expected weaker effect associated with each rabbi.

Our experimental protocol specifies that we reject the Null hypothesis for any rabbi having a p-value less than or equal to .25. Any rabbi for which we reject the Null hypothesis is considered to be encoded. Under the Null hypothesis, the expected number of rabbis having an observed p-value of less than .25 out of the 31 rabbis of the combined list is 31 x .25 = 7.75.

The experimental results show that for the combined list there were 17 rabbis having an experimental p-value less than or equal to .25. The rabbis having a significantly more compact arrangement than expected by chance are Rabbis Abraham of Narbonne, Eliezer Ashkenzai, David Oppenheim, David Nieto, Chaim Abulafia, Chaim Capusi, Chaim Shabetai, Yehuda Chasid, Yoseph Teomin, Yaakov Beirav, The Marahal, The Yaabetz, Menachem Krochmal, Moshe Zacuto, Immanuel Hai Ricchi, Shalom Sharabi, and Meir Eisenstat .

How unexpected is it to observe 17 rabbis with p-values less than .25 out of a set of 31 rabbis? Under the Null hypothesis, the probability that any p-value will be less than or equal to .25 is .25. We may reasonably assume that the experimentally observed p-values are independent because the result for each rabbi was made up of many appellation date pairs no one of which was common to more than one rabbi. We made 31 experiments and observed 17 successes. Under these conditions the number of successes is binomially distributed. The probability of observing 17 or more successes is 3.67×10^{-4}. Therefore, at the 10^{-3} significance level, the Null hypothesis must be rejected and we conclude that under the Null hypothesis it would be very unlikely to observe 17 rabbis out of 31 each producing a p-value of .25 or less.

Our results for testing the Null hypothesis against alternative (4) are robust to the choice of a .25 p-value threshold. Indeed we must reject the Null hypothesis at the significance level of .001 for all p-value thresholds between .12 and .38.

For comparison we also examine all the lists against both texts. The table of p-values is shown below. At the .001 signficance level, we must reject the Null hypothesis for all

lists but the McKay list on the *Genesis* text. And for each of the lists we cannot reject the Null hypothesis on the *War and Peace* text. Also notice that as with the test of the Null hypothesis against the first alternative hypothesis, the combined list has a smaller p-value (stronger encoding effect) compared to the WRR list and the McKay list.

List	***Text***	***# Successful rabbis***	***Total # rabbis***	***P-value***
WRR List One	*Genesis*	17	32	6.00×10^{-4}
WRR List Two	*Genesis*	16	30	8.19×10^{-4}
McKay List	*Genesis*	13	30	2.16×10^{-2}
Combined List	*Genesis*	17	31	3.67×10^{-4}
WRR List One	*War and Peace*	7	32	7.22×10^{-1}
WRR List Two	*War and Peace*	10	30	1.97×10^{-1}
McKay List	*War and Peace*	15	30	2.75×10^{-3}
Combined List	*War and Peace*	14	31	1.15×10^{-2}

Table summarizing the p-values associated with testing the Null hypothesis against the alternative that there are more rabbis having small compactnesses than expected by chance. Small compactness here means that the p-value for the rabbi is less than .25. Notice that the Null hypothesis of no Torah code effect cannot be rejected for the McKay list for both the Genesis text and the War and Peace text while the Null hypothesis of no Torah code effect must be rejected for the WRR and combined lists for the Genesis text. Also notice that p-value of the combined list for the War and Peace text is weaker than the p-value of the McKay list for the War and Peace text.

The McKay Study

In the light of our study, we are in a position to critically scrutinize the arguments made by McKay et. al. First they argue that the permutation test of WRR is unsatisfactory because the number of (appellation, date) pairs in each permutation changes. Our methodology uses a control text population. The (appellation, date) pairs stay exactly the same for each trial of our experiment and the number of terms in the geometric mean stays exactly the same for each trial of our experiment.

Second they argue that the WRR method in multiplying the c-values becomes overly sensitive to the values of the smallest c-values. In our case, we multiply the rank normalized compactness values of the appellation date pairs. The rank normalization non-linearly maps the smallest compactness value to one half divided by the number of sampled texts, if the smallest compactness value is unique among the sampled texts. The number of terms in the product is the same for each trial and the multiplication is therefore equivalent to a geometric mean. In a test of the Null hypothesis against an alternative hypothesis which specifies that there are more terms than expected by chance that are smaller than expected by chance, it is exactly the effect of the small valued terms that must be brought out. Either the statistical methodology tries to select and count the number of small terms as it does in the Best Star Team statistic or it tries to recover the degree to which there are more than expected smaller valued terms than expected in an overall way without any selection as in the geometric mean, as we do in our test of the

Null hypothesis against alternative hypothesis (2). In general, when there is not a well separated group of small values, statistical methods that would make the selection are not as powerful, in terms of false alarm and misdetect rates, as methods that gather an overall statistic such as a geometric mean. Furthermore, the geometric mean is more powerful, in terms of false alarm and misdetect rates than the arithmetic mean for these purposes. High sensitivity to small values is exactly is what is required for a good detector in these circumstance. The McKay et. al. argument in this instance is just completely off base statistically.

Thirdly, they argue that the choice of date form selected by WRR is advantageous to producing a small p-value over other forms which yield p-values that are not particularly significant. This argument is unbelievable. A new experiment is not done without a history of experiments that lead up to the choices made in the new experiment. On the basis of previous experiments, those prior to WRR List One as well as other perhaps informal experiments WRR did, they must have noticed that the date form they used worked more often than other forms. There is no hypothesis that says all date forms are encoded. There was no scientific or statistical requirement that WRR should use all date forms, when there is already prior evidence for using the form which the hypothesized encoding uses. Under these conditions WRR did what any reasonable researcher would also do: Keep to a protocol having components that worked previously.

Fourthly, they argue that WRR used far less than half of all the appellations by which their rabbis were known. Had they used all the appellations there would not have been any interesting results. We did not have access to all the appellations that McKay et. al. think the rabbis had. We did have access to the 31 appellations that McKay et. al. added to those of WRR. And our experiment shows a smaller or equal p-value with the combined list than the WRR List, indicating that there were some encodings missed by the WRR protocol. So to the extent possible with the information we had, we demonstrated that the inclusion of missing appellations made the result stronger in the *Genesis* text and not weaker as assumed by McKay et. al. Furthermore, the result in the *War and Peace* text became weaker using the combined list and indeed we were not able to reject the Null hypothesis for *War and Peace* at the .001 significance level in the rabbi by rabbi experiments (alternative hypotheses (3) and (4)) or against alternative hypothesis (2). The fact that there was statistical significance in the test of the Null hypothesis against alternative (1) suggests that the statistical methodology used for this test has a greater false alarm rate than the other tests.

Fifthly they argue by a cooked demonstration that by tuning the selected appellations and their spellings to a text such as *War and Peace*, they could produce a p-value that was significant for *War and Peace*. Indeed, with our improved compactness measures and analysis protocols, we did observe that the p-value for the McKay list in *War and Peace* was significant at the .001 significance level against alternative hypotheses (1), (2) and (3) and as well it was significant in the *Genesis* text against alternative hypothesis (1). But because the behavior is different for the H_1 and H_2 measures between the *Genesis* text and the *War and Peace* text against alternative hypothesis (1) we hypothesize that the way the McKay list was cooked was in adding what are most probably incorrect

appellations that had very good H_2 encodings for one pair of ELSs for the corresponding pair of (appellation, dates). The real encodings most often do not appear in this way. They appear moderately encoded with ELS encoding redundancy. We believe that this difference is the cause of the observed behavior. A future study is planned on this issue.

Sixthly they argue that the WRR *Omega* compactness measure was tuned to the data. We can now provide evidence that this argument has to be fallacious. With our improved protocol, the WRR *Omega* measure was among the poorer less powerful measures and was ruled out of use by our earlier experiments on WRR rabbis list one. Furthermore, our compactness measure worked as expected for the WRR List One, the McKay List, and the Combined List on the *Genesis* text and the *War and Peace* text. It worked in a consistent way with the lists taken as a whole and with the lists divided up rabbi by rabbi. This kind of behavior is not due to chance variations that can be capitalized by tuning. It is actually capturing and measuring a real effect.

Concluding Discussion

In this note we explored the correctness of the McKay et. al. argument that if a more complete appellation list had been used instead of the WRR list, the observed Torah code effect in the *Genesis* text would disappear, because the Null hypothesis of no Torah code effect is true. Therefore, we performed an improved set of Torah code experiments using a more complete list formed by combining the WRR and McKay appellation lists. We used an ELS random placement control text population and some compactness measures with better statistical properties than the original WRR compactness measure. We employed a better statistical analysis methodology. We tested the Null hypothesis of no Torah code effect against four different alternative hypotheses:

- the observed compactnesses of the appellation date pairs among all the rabbis is smaller than expected by chance;
- the number of appellation date pairs having small values among all the rabbis is larger than expected by chance;
- the observed p-values of the rabbis is smaller than expected by chance;
- the number of rabbis having a small p-value is larger than expected by chance.

At the .001 significance level, we had to reject the Null hypothesis against each of the four alternatives. Therefore, contrary to McKay et. al. we conclude that with a more complete appellation list, the Torah code effect did not go away. In fact we found a slightly stronger effect with the combined appellation list than with either the McKay list or the WRR list. On tracing to what additional appellations from the McKay list this slightly stronger effect occurs in the combined list, we found the biggest contributors were the appellations for Rabbi Chaim Abulafia: המהרחא, מהרחא. The table below shows the p-level for Rabbi Chaim Abulafia just due to these two appellations.

Compactness Measure	**H_1**	**H_2**	**H_3**	**H_4**	**Exp.**
P-value	.000250	.443850	.010450	.412650	.0008

The p-value associated with $\min(H_1,H_2,H_3,H_4)$ for the WRR appellations of Rabbi Chaim Abulafia is .01. The two appellations of McKay decrease this p-value by more than a factor of 10.

We logically conclude that for the great rabbis experiment there is a Torah code effect and that the observed effect is not due to a small number of very compact terms or to a small number of rabbis having some of their appellation dates encoded. Rather the effect is distributed among many appellation date pairs and among many of the rabbis.

Starting from the point that the Torah code effect is real for the great rabbis experiment, we see that there is a significant difference between the test of the Null hypothesis of no Torah code effect against each of the four different alternatives. Our results indicate that for the statistical tests employed, testing the Null hypothesis against hypothesis alternatives (3) and (4), dealing with the rabbi by rabbi results, seem to have better false alarm rates than the tests against alternative hypotheses (1) and (2), which put all ELS appellation date pairs into one pile, so to speak. This may suggest that a better statistical test is needed for determining whether a subset of appellation date pairs are encoded. This is undoubtedly the reason why the cooked McKay list has a smaller p-value than we would like on the War and Peace text. Our future work will be directed towards developing a better statistical detector with a smaller false alarm rate..

References

M. Bar-Hillel, Dror Bar-Natan, and Brendan McKay, "Torah Codes: Puzzle and Solution, *Chance*, Vol 11, pp.13-19.

Randall Ingermanson, *Who Wrote the Bible Code*, Waterbrook Press, Colorado, 1999.

Brendan McKay, Dror Bar-Natan, Maya Bar-Hillel, and Gil Kalai, "Solving the Bible Code Puzzle", *Statistical Science*, Vol 14, No 5, May 1999, pp. 150-173 .

Doron Witztum, Eliyahu Rips, and Yoav Rosenberg, ``Equidistant Letter Sequences in the Book of Genesis," *Statistical Science*, Vol. 9, No. 3, 1994, pp. 429-438.

Doron Witztum, *New Statistical Evidence for a Genuine Code in Genesis,* www.torahcode.il, 2000.

Doron Witztum, *Of Science and Parody: A Complete Refutation of MBBK's Central Claim*, http://www.torahcodes.col.il.paro_hb.htm, (2000).

Dror Bar-Natan, *Equidistant Letter Sequences in Tolstoy's "War and Peace"*, http://www.math.toronto.edu/~drorbn/Codes/WNP/main_bh.html, 1997.