SOLVING THE BIBLE CODE PUZZLE

BRENDAN MCKAY, DROR BAR-NATAN, MAYA BAR-HILLEL, AND GIL KALAI

ABSTRACT. A paper of Witztum, Rips and Rosenberg in this journal in 1994 made the extraordinary claim that the Hebrew text of the Book of Genesis encodes events which did not occur until millennia after the text was written. In reply, we argue that Witztum, Rips and Rosenberg's case is fatally defective, indeed that their result merely reflects on the choices made in designing their experiment and collecting the data for it. We present extensive evidence in support of that conclusion. We also report on many new experiments of our own, all of which failed to detect the alleged phenomenon.

CONTENTS

Date: June 1999. In press for **Statistical Science**, May 1999 issue.

1. Introduction

Whilst history records a great many claims of sacred texts hiding messages or meanings beyond their manifest content, it seems that only in the past century have serious efforts been made to prove the existence of such messages by scientific means. Examples include the Christian scriptures (Panin, 1908; McCormack, 1923) and the Islamic scriptures (Khalifa, 1992). However, although those "discoveries" might appear astonishing at first glance, a modest amount of effort is sufficient to expose the invalid statistics (and, all too often, sleight of hand) beneath the thin façade of "science" (McKay, 1999a).

A recent paper of Witztum, Rips and Rosenberg (1994), whom we will refer to as WRR, is not obviously in the same category. Instead, it has the form of a carefully designed and executed experiment. Our purpose here is to see whether this apparent solidity holds up under thorough scrutiny. WRR's paper (1994) is the main focus of this paper; we will refer to it as WRR94.

WRR claim to have discovered a subtext of the Hebrew text of the Book of Genesis, formed by letters taken with uniform spacing. Their paper was reprinted in full in a book of Drosnin (1997) that has been a best-seller in many languages, so it is possibly the most printed scientific paper of all time. It has spawned a large "Bible codes" industry, with at least eight books and three television documentaries so far and a movie in production. People wishing to find "codes" for themselves have the choice of many commercially available programs. Several large religious organizations (Jewish and Christian) have adopted the "codes" as part of their repertoire. Thus, even though WRR94 did not attract much previous scientific attention, it is clearly in the public interest to examine the evidence in detail.

Consider a *text*, consisting of a string of letters $G = g_1 g_2 \cdots g_L$ of length L, without any spaces or punctuation marks. An *equidistant letter sequence (ELS)* of length k is a subsequence $g_n g_{n+d} \cdots g_{n+(k-1)d}$, where $1 \le n, n+(k-1)d \le L$. The quantity d, called the *skip*, can be positive or negative.

As one would expect, an ELS will sometimes spell out a meaningful word. WRR's work was motivated by their informal observation that when the Hebrew text of Genesis is written as a string around a cylinder with a fixed circumference, they often found ELSs for two thematically or contextually related words in physical proximity. To illustrate the concept, we give an English example. In Figure 1 we show an 8×18 rectangle cut from the Manifesto (Kaczynski, 1995) written by the "Unabomber," Ted Kaczynski, when its text is written around a cylinder with circumference of 158 letters. ELSs for the words "mail" and "bombs" are seen to appear close together. Readers are invited to find the slogan "Free Ted!" also hidden in the picture.

N	D	W	I	L	L	D	I	S	C	U	S	S	I	T	L	A	T
P	R	O	$\boxed{B}$	L	E	M	S	A	T	F	I	R	S	T	H	E	W
S	T	E	N	D	T	$\boxed{O}$	B	E	C	O	M	E	D	E	C	A	D
U	A	L	L	Y	B	E	C	O	$\boxed{M}$	E	B	O	R	E	D	H	E
T	H	I	N	G	E	L	S	E	T	O	O	$\boxed{B}$	T	A	I	N	T
U	T	E	F	F	O	R	T	H	E	N	C	E	H	I	$\boxed{S}$	B	O
C	O	$\boxed{M}$	P	$\boxed{A}$	T	$\boxed{I}$	B	$\boxed{L}$	E	W	I	T	H	S	U	R	V
M	A	N	B	E	I	N	G	N	E	E	D	S	G	O	A	L	S

FIGURE 1. Messages in the Unabomber Manifesto

Many more examples of such letter arrays have been presented by Drosnin (1997), Satinover (1997), Witztum (1989) and Young (1997), for the Bible, or by McKay (1999b) and Thomas (1997), for other texts. It is acknowledged by WRR that they can be found in any sufficiently long text. The question is whether, as WRR claim, the Bible contains them in compact formations more often than expected by chance.

In WRR94, WRR presented what they called a "uniform and objective" list of word pairs—names and dates of birth or death of famous rabbis from Jewish history—and analyzed their proximity as ELSs in a formal sense inspired by the informal observations described above. The result, they claimed, is that the proximities are on the whole much better than expected by chance, at a significance level of 1 in 60,000. Since the word pairs refer to people who lived millennia after the book of Genesis was written, one can only describe the conclusion as astonishing.

This paper scrutinizes almost every aspect of the alleged result. After a brief exposition of WRR's work in Sections 2 and 3, we demonstrate in Section 4 that WRR's method for calculating significance has serious flaws. In Section 5 we question the quality of WRR's data. Most importantly, we show that the data was very far from tightly defined by the rules of their experiment. Rather, there was enormous "wiggle room" available, especially in the choice of names for the famous rabbis. The literature contains a considerable number of variations in names and their spellings, as well as other appellations such as nicknames and acronyms, but WRR used only a fraction of them. WRR also had substantial choice in other aspects of the experiment, including the method of analysis.

It is valid to raise the question of whether this lack of tightness in the design of the experiment is at the heart of the result. In precise terms, we ask two questions:

- Was there enough freedom available in the conduct of the experiment that a small significance level could have been obtained merely by exploiting it?
- Is there any evidence for that exploitation?

The first question is answered affirmatively in Section 6, where we employ a small part of the same freedom to construct an alternative data set that appears to produce an equally small significance level using the text of *War and Peace* instead of the text of Genesis. To answer the second question, in Section 7 we examine a very large number of minor variations on WRR's experiment and find that the result becomes weaker in the great majority of cases. This appears very unlikely to have occurred by chance, suggesting that WRR's data suffers from systematic bias. This theory is supported in Section 8, which shows that WRR's data also matches common naive statistical expectations to an extent unlikely to be accidental.

In Sections 9 and 10, we discuss other ELS experiments. We report that the other experiments claimed to have detected "codes" suffer from the same problems as beset the experiment in WRR94. In contrast, all of our own experiments failed to find any trace of a non-chance ELS phenomenon. Finally, in Section 11 we describe what is known about the history of the text of Genesis, and conclude that no "codes" in the original text could have survived the long process of textual transmission from the original edition to what we have today.

Nontechnical popular expositions of some of this work have previously been published by Bar-Hillel, Bar-Natan and McKay (1997, 1998). Even in the present paper, the reader may safely skip over the more technical sections and still gain a fair appreciation of our study.

Over the course of our long investigation, we have studied many more aspects of the subject than we are able to present here. Nothing we have chosen to omit tells a story contrary to the story here.

Much further information on this subject, including coverage of the argument engendered by this paper, can be found on McKay's web site (1999b). Other informed articles were authored by Perakh (1998), Simon (1998) and Tigay (1998).

2. Overall closeness and the permutation test

The work of WRR is based on a very complicated function $c(w, w')$ that measures some sort of proximity between two words w and w', according to the placement of their ELSs in the text. A precise definition is given in Appendix A, but the details are only needed for the more technical aspects of Section 7. Here we will describe how WRR used $c(w, w')$

to define an aggregate measure of closeness for a set of word pairs and how that aggregate measure was in turn used to compute a "significance level".

As the details in Appendix A explain, $c(w,w')$ is sometimes undefined for a word pair (w,w'), and is otherwise a nonzero number in $[0,1]$. Ignoring undefined values altogether, suppose $c_1, c_2, \ldots, c_N$ is the sequence of $c(w,w')$ values for some sequence of N word pairs. WRR use two methods of turning this sequence of values into a single value. Let X be the product of the c_i's, and m be the number of them which are less than or equal to 0.2. Define

$$P_1 = \sum_{i=m}^{N} \binom{N}{i} \left(\frac{1}{5}\right)^i \left(\frac{4}{5}\right)^{N-i},$$

$$P_2 = X \sum_{i=0}^{N-1} (-1)^i (\log X)^i / i!\ .$$

The rationale for P_1 and P_2, as stated by WRR (1994), is that they would have simple meanings if the c_i's were independent uniform variates in $[0,1]$. Namely, P_1 would be the probability that the number of values at most 0.2 is m or greater, and P_2 would be the probability that the product is X or less. Neither independence nor uniformity hold in this case, but WRR claim that they are not assuming those properties. They merely regard P_1 and P_2 as arbitrary indicators of aggregate closeness.

The paper WRR94 considers a data set consisting of two sequences W_i and W'_i $(1 \le i \le n)$, where each W_i and each W'_i are possibly-empty sets of words. The *permutation test* defined there is intended to measure if, according to the distance measures P_1 and P_2, the words in W_i tend to be closer to the words in W'_i than expected by chance, for all i considered together. It does this by pitting distances between W_i and W'_i against distances between W_i and W'_j, where j is not necessarily equal to i.

Let π be any permutation of $\{1, 2, \ldots, n\}$, and let π_0 be the identity permutation. Define $P_1(\pi)$ to be the value of P_1 calculated from all the defined distances $c(w,w')$ where $w \in W_i$ and $w' \in W'_{\pi(i)}$ for some i. Then the *permutation rank of* P_1 is the fraction of all $n!$ permutations π such that $P_1(\pi)$ is less than or equal to $P_1(\pi_0)$. Similarly for P_2. We can estimate permutation ranks by sampling with a large number of random permutations.

3. The Famous Rabbis experiment

The experiment in WRR94 involves various appellations (names, nicknames, acronyms, etc.) of famous rabbis from Jewish history paired with their dates of death and, where available, birth. (Dates in Hebrew are written using letters only, without numerals.)

Interpretation of some of our observations reported below depends on the details of the chronology of the experiment. Since much of it is contentious and of considerable public interest, we provide what we believe to be an accurate account of as much of the history as can be established from the documentary evidence.

1. The idea of using the names and dates of famous rabbis was conceived about 1985. WRR claim that the first-ever experiment performed was on a set of 34 rabbis, together with appellations and dates, identical in every way to Table 1 of WRR94, and that they had no prior knowledge of rabbis having their names appear close to their dates as ELSs. However, an early lecture of Rips (1985) described an experiment with a particular subset of "19 or 20" rabbis. Be that as it may, the list of appellations and dates of the 34 rabbis, and a definition of $c(w, w')$ apparently consistent with that later defined in WRR94, appeared in a preprint of WRR (1986). The definition of P_2 also appeared there, together with what we will call the P_1-precursor: the number of $c(w, w')$ values less than or equal to 0.2, expressed as a number of standard deviations above the expected value, assuming a binomial distribution. The value of P_2 and, implicitly, the value of P_1, were presented as probabilities, in disregard of the requirements of independence and uniformity of the $c(w, w')$ values that are essential for such an interpretation.
2. At some point the work was brought to the attention of Persi Diaconis, then Professor of Statistics at Harvard University, who requested that a standard statistical test be used to compare the distances against those obtained after permuting the dates by a "randomly chosen cyclic shift" (Diaconis, 1986). He also requested "a fresh experiment on fresh famous people". In 1987 a second preprint (WRR, 1987) appeared, containing the list of 32 rabbis which appear in Table 2 of WRR94, which WRR had produced as a second sample. That preprint contained the distances for the new sample, and also for a cyclic shift of the dates (not random as Diaconis had requested, but matching rabbi i to date $i + 1$) after certain appellations (those of the form "Rabbi X") were removed. The requested significance test was not reported; instead, the statistics P_2 and, again implicitly, P_1 were once again incorrectly presented as probabilities. There was still no permutation test at this stage, except for the use of a single permutation.
3. About 1988, a shortened version of WRR's preprint (1987) was submitted to a journal (Proceedings of the National Academy of Sciences of the USA) for possible publication. To correct the error in treating P_{1-4} (that is, P_1, P_2, P_3 and P_4) as probabilities, Diaconis proposed a method that involved permuting the columns of a 32×32 matrix, whose (i, j)th entry was a single value representing some sort of aggregate distance between all the appellations of rabbi i and all the dates of rabbi j. This proposal

was apparently first made in a letter of May 1990 to the Academy member handling the paper, Robert Aumann, though a related proposal had been made by Diaconis in 1988. The same design was again described by Diaconis in September (Diaconis, 1990), and there appeared to be an agreement on the matter. However, unnoticed by Diaconis, WRR performed the different permutation test described in Section 2. A request for a third sample, made by Diaconis at the same time, was refused.

4. After some considerable argument, the paper was rejected by the Proceedings of the National Academy of Sciences and sent instead to Statistical Science in a revised form that only presented the results from the second list of rabbis. It appeared there in 1994, without commentary except for the introduction from editor Robert Kass: "Our referees were baffled . . . The paper is thus offered . . . as a challenging puzzle." (Kass, 1994; cf. Kass, 1998).

In the experiment presented in WRR94, the word set W_i consists of several (from 1 to 11) appellations of rabbi i, and the word set W'_i consists of several ways of writing his date of birth or death (from 0 to 6 ways per date), for each i. As mentioned in (2) above, WRR also used data modified by deleting the appellations of the form "Rabbi X". We will follow WRR in referring to the P_1 and P_2 values of this reduced list as P_3 and P_4, respectively. The unreduced list produces about 300 word pairs, of which somewhat more than half give defined $c(w, w')$ values.

The permutation ranks estimated for P_2 and P_4 were 5×10^{-6} and 4×10^{-6}, respectively, and about 100 times larger (i.e., weaker) for P_1 and P_3. The oft-quoted figure of 1 in 60,000 comes from multiplying the smallest permutation rank of P_{1-4} by 4, in accordance with the Bonferroni inequality. These permutation ranks estimates are in fact too large, perhaps due to the sampling error caused by using only one million random permutations. Both WRR (1995) and ourselves obtain even more impressive values if we compute them more accurately. Using 200 million random permutations, we estimate the permutation ranks for P_2 and P_4 to be about 1.9×10^{-6} and 6.8×10^{-7}, respectively.

WRR's first list of rabbis and their appellations and dates appeared in WRR94 too, but no results are given except some histograms of $c(w, w')$ values. Since WRR have consistently maintained that their experiment with the first list was performed just as properly as their experiment with the second list, we will investigate both.

4. Critique of the test method

A critique of WRR's test method from several points of view is given by Hasofer (1998). We will not repeat those points here, except to note that Hasofer demonstrates their test statistic to be fraught with anomalies, such as sometimes being small when we expect it to be large. He also criticizes WRR's failure to present an explicit alternative hypothesis. Readers should consult his paper for the details.

WRR's null hypothesis H_0 has some difficulties. As defined in WRR94, H_0 says that the permutation rank of each of the statistics P_{1-4} has a discrete uniform distribution in $[0, 1]$. It is worth considering whether that null hypothesis makes sense and whether its rejection has the implications that are commonly claimed.

If there is no prior expectation of a statistical relationship between the names and the dates, we can say that all permutations of the dates are on equal initial footing and therefore that the null hypothesis holds on the assumption of "no codes". However, the test is unsatisfactory for the following reason: even though WRR claim to be detecting a property of the text of Genesis, the distribution of the permutation rank *conditioned on the list of word pairs*, is not uniform at all. We show this below. Because of this property, rejection of the null hypothesis may say more about the word list than about the text.

To see that WRR's null hypothesis does not hold conditional on the list of word pairs, we need to look at the mathematics of the distance function $c(w, w')$. The distribution of $c(w, w')$ for random words w and w', and fixed text, is approximately uniform. However, any two such distances are dependent as random variables. The most obvious example of dependence (of many that are present) is between $c(w, w')$ and $c(w, w'')$, where there is an argument w in common, because both depend on the number and placement of the ELSs of w. Because presence of such dependencies amongst the distances from which P_2 is calculated changes the *a priori* distribution of P_2, and because this effect varies for different permutations, the *a priori* rank order of the identity permutation is not uniformly distributed.

An analogy might make the difficulty clearer. The performance of athletes in the long jump can be greatly affected by the strength of the wind, especially on a windy day. If we think of the competition as based on the premise "we are giving the athletes the same chance of winning", the test is fair because each athlete has the same chance of being hindered or assisted by the wind. However, the winner might be determined by the wind, rather than by the athletes' skills. We consider this unsatisfactory because the premise we *really* want to base the competition on is "the chance of winning depends only on skill". However, the unpredictable nature of the wind invalidates this premise. In the same way, the result of WRR's permutation test may reflect (at least to some extent) uninteresting properties of the word list rather than an extraordinary property of Genesis.

The result of the dependence between $c(w, w')$ values is that the *a priori* distribution of $P_2(\pi)$, given the word pairs, rests on such mundane matters as the number of word pairs that π provides (just as, in our analogy, the chance of each person winning depends on the wind strength). Since different permutations provide different numbers of word pairs (due to the differing sizes of the sets W_i and W_i'), they do not have an equal chance of producing the best P_2 score. It turns out that, for the experiment in WRR94 (second list), the identity permutation π_0 produces more pairs (w, w') than about 98% of all permutations. The effect of this extremeness on the result is hard to pin down but, whatever it is, we certainly cannot attribute it to the text for the simple reason that it is completely independent of the text. In fact, the number of word pairs is only one example of text-independent asymmetry between different permutations. Other examples include differences with regard to word length and letter frequency.

These concerns do not apply, or are greatly reduced, for the method proposed by Diaconis (1990). For the record, the most obvious definition of his 32×32 matrix (using the average distance), and the definition he informally used himself (using the minimum distance), both produce results hundreds of times weaker than WRR obtained using their own method.

Serious as these problems might be, we cannot establish that they constitute an adequate "explanation" of WRR's result. For the sake of the argument, we are prepared to join them in rejecting their null hypothesis and concluding "something interesting is going on". Where we differ is in what we believe that "something" is.

Sensitivity to a small part of the data.

A worrisome aspect of WRR's method is its reliance on multiplication of small numbers. The values of P_2 and P_4 are highly sensitive to the values of the few smallest distances, and this problem is exacerbated by the positive correlation between $c(w, w')$ values.

Due in part to this property, WRR's result relies heavily on only a small part of their data. We will illustrate this with two observations about the experiment in WRR94, using the method of analysis employed there.

- If the 4 rabbis (out of 32) who contribute the most strongly to the result are removed, the overall "significance level" jumps from 1 in 60,000 to an uninteresting 1 in 30. Historically speaking, these rabbis are not particularly important compared to the others.
- One appellation (out of 102) is so influential that it contributes a factor of 10 to the result by itself. Removing the five most influential appellations hurts the result by a factor of 860. Again, these appellations are not more common or more important than others in the list in any previously recognized sense.

It should be obvious from these facts that a small change in the data definition (or in the judgement or diligence of the data collector) might have a dramatic effect. More generally, the result of the experiment is extraordinarily sensitive to many apparently minor aspects of the experiment design, as we will amply demonstrate.

These properties of the experiment make it exceptionally susceptible to systematic bias. As we shall see, there appears to be good reason for this concern.

5. Critique of the list of word pairs

The image presented by WRR of an experiment whose design was tight and whose implementation was objective falls apart upon close examination. We will consider each aspect of the data in turn.

The choice of rabbis.

The criteria for inclusion of a rabbi in WRR's lists were quite mechanical. They were taken from Margaliot's *Encyclopedia of Great Men of Israel* (1962). For the first list, the rabbi's entry had to be at least 3 columns long and mention a date of birth or death. For the second list, the entry had to be from 1.5 columns to 3 columns long. However, these mechanical rules were carried out in a careless manner. At least seven errors of selection were made: in each list there are rabbis missing and rabbis who are present but should not be. However, these errors have a comparatively minor effect on the results.

The choice of dates.

Each rabbi has potentially two dates, one of birth and one of death, though in most cases Margaliot (1962) only lists the death date. In WRR94 we read "our sample was built from a list of personalities ... and the dates ... of their death or birth. The personalities were taken from [Margaliot]", and readers can be forgiven for inferring that the dates came from there also. However, from WRR's preprints (1986, 1987), we know that they came from a wide variety of sources. Some dates given by Margaliot were omitted on the grounds that they are subject to dispute, but at least two disputed dates were kept. Other dates were changed in favour of sources claimed to be more authoritative than Margaliot, but at least two probably wrong dates were not corrected. One date which was neither a date given by Margaliot nor a correction of one was introduced from another source. However, several other dates readily available in the literature were not introduced. The details appear in Appendix B.

The choice of date forms.

In addition to choosing which dates to use, there was a choice of how to write the dates. Only the day and month were used, not the year. Particular names (or spellings) for the months of the Hebrew calendar were used in preference to others, and the standard practice of specifying dates by special days such as religious holidays (used in WRR's main source Margaliot (1962), for example) was avoided.

To write the day and the month, WRR used three forms, approximately corresponding to the English forms "May 1st," "1st of May" and "on May 1st". They did not use the obvious "on 1st of May," which is frequently used by Margaliot, nor any of a number of other reasonable ways of writing dates (details below). Most surprising is how they wrote the fifteenth and sixteenth of each month. These are customarily written using the letters representing 9+6 (or 9+7), avoiding the letter pairs representing 10+5 (or 10+6) for religious reasons. The nonstandard forms were in occasional use centuries ago, but are now so obscure that few except scholars have seen them used. Despite this, WRR chose to use both—a choice greatly in their favour, as we shall see in Section 7.

At least five additional date forms are used in Hebrew in addition to the three WRR used, so it is interesting so see how they perform. In Table 1 we show what happens if each date form is used by itself, using this key: D is the day of the month, M is the month, b is the prefix "in," ℓ is the prefix "of" and *shel* is the word "of". We also give, in the last row of the table, what happens if the dates are written in precisely the form in which they appear in WRR's source encyclopedia (Margaliot, 1962). The values given are permutation ranks.

Date form	Used by WRR?	List 1	List 2
D M	yes	0.165751	0.000017
bD M	yes	0.000008	0.008844
D bM	yes	0.006070	0.008804
bD bM	no	0.068478	0.429256
D ℓM	no	0.581777	0.274167
bD ℓM	no	0.281509	0.618128
D *shel* M	no	0.711538	0.046468
bD *shel* M	no	0.467761	0.135884
Margaliot	partly	0.070780	0.277658

TABLE 1. Different date forms used alone. (Least of P_{1-4} permutation ranks)

The lesson to take from Table 1 is that each of the three forms used by WRR perform very well in one or both lists, but the other forms are failures. More information on this subject appears in Appendix B.

The choice of appellations.

We now come to the most serious problem with the data: the choice of appellations to use for the famous rabbis. The rabbinical literature abounds with such appellations, often with multiple variations in spelling and use of articles. An example in English will illustrate what an "appellation" is in this context. A certain celebrated person can be referred to as

John F. Kennedy, Jack Kennedy, JFK, Mr. President, Mr. J. Kennedy, Kennedy, or "the man who accompanied Jackie to Paris," to list but a few. Similarly, many famous rabbis of history can be, and are, referred to in a considerable number of ways. Acronyms and other abbreviations are especially common. (For example, "Rambam" is an acronym for Rabbenu Moshe ben Maimon, also known as Maimonides.)

Since WRR used far less than half of all the appellations by which their rabbis were known, the issue of how the selection was made is central to the interpretation of their experiment. Their paper WRR94 has only this to say on the issue: "The list of appellations for each personality was prepared by Professor S. Z. Havlin, of the Department of Bibliography and Librarianship at Bar-Ilan University, on the basis of a computer search of the 'Responsa' database at that university." This has led to a widely held misconception that the list was comprehensive or that the selection was rigorous and mechanical. Not so. Many of the appellations in Responsa do not appear in WRR94 and vice versa. Moreover, Menachem Cohen of the Department of Bible at Bar-Ilan University, after studying WRR's lists, reported that they have "no scientific basis, and [are] entirely the result of inconsistent and arbitrary choice" (Cohen, 1997a).

The earliest available documents on the experiment (Rips, 1985; Witztum, 1989; WRR, 1986; WRR, 1987) do not state that the lists of appellations were prepared by an independent source. Rips did not mention Havlin at all in his early lecture (1985), but described appellation selection differently:

> *There may be various ways of writing a name. We took every possible variation we could think of. For instance, Ha'Gaon... or Eliyahu... or, say, Rabbi Eliyahu. If any additional variation comes to mind, we must include it. We simply took every possible variant that we considered reasonable.* [Ellipses in original.]

Havlin is acknowledged in WRR's first two preprints (1986, 1987), but only for providing "valuable advices" [sic]. The earliest clear claim we could uncover that the appellations were Havlin's work was in preprints of WRR94 from about 1992. Details were provided years later by Havlin himself (1996), who certified explicitly that he had prepared the lists on his own, and gave explanations for many of his decisions. He acknowledged making several mistakes, not always remembering his reasoning, and exercising discretionary judgement based on his scholarly intuition. He also admitted that if he were to prepare the lists again, he might decide differently here and there.

The question has to be asked whether the strong result in WRR94 might be largely attributable to a biasing of the appellation selection, fortuitously or otherwise, towards those performing well in WRR's experiment. The most immediate issue is whether such biasing is technically feasible. Was the flexibility available in the selection of appellations at the time the lists were prepared sufficient that biased selection could produce a strong result? The great sensitivity of WRR's result to the data that we demonstrated in Section 4 suggests that the "wiggle room" is more than enough. In the next section, we will demonstrate that this intuition is correct.

6. Appellations for *War and Peace*

An Internet publication by two ofthe present authors (Bar-Natan and McKay, 1998), presented a new list of appellations for the 32 rabbis of WRR's second list. The appellations are not greatly different from WRR's: 83 were kept, 20 were deleted and 29 additional appellations were added. Many of the changes were simply replacements of one valid spelling by another. The punch line is that the new set of appellations produces a "significance level" of one in a million when tested in the initial 78,064 letters (the length of Genesis) of a Hebrew translation of Tolstoy's *War and Peace*, and produces an uninteresting result in Genesis. Exactly the same text of *War and Peace* is used for control tests in WRR94.

All of our changes were justified either by merely being correct, or by virtue of being no more doubtful than some analogous choice made in WRR's list. For example, whereas WRR used one common Hebrew spelling of the name "Horowitz," we used a different common spelling. When they omitted one common appellation, we inserted it and deleted another. And so on. Our list of appellations does not aspire to be perfect, merely to be of quality commensurate with that of WRR's list. As verified by Menachem Cohen, there is "no essential difference" between WRR's list and ours (Cohen, 1997a). (Amusingly, one knowledgable rabbi who inspected both lists pronounced them "equally appalling".)

This demonstration demolishes the common perception and oft-repeated claim that the freedom of movement left by the rules established for WRR's first list was insufficient by itself to explain an astounding result for the second list.

The appellation list of Bar-Natan and McKay (1998) has been the subject of concerted attack (Witztum 1998a). The essence of his thesis is that WRR's lists were governed by rules, and that the changes made in the second list to tune it to *War and Peace* violate these rules. However, most of these "rules" were only laid out nine to ten years after WRR's two lists were composed, in a lengthy letter written by Havlin (1996) in response to some questions we raised, and had never been publicly mentioned before. While the letter offers many explanations and examples of Havlin's considerations when selecting among possible appellations, they are far from being rules, and are fraught with inconsistency. Moreover, when rules for a list are laid out a decade after the lists, it is not clear whether the rules dictated the list selections, or just rationalize them. Besides, as Bar-Natan and McKay amply demonstrate (1999), these "rules" were inconsistently obeyed by WRR.

Most of Witztum's criticisms are inaccurate or mutually inconsistent, as the following two examples illustrate:

1. Witztum argues against our inclusion of some appellations on the grounds that they are unusual, yet defends the use in WRR94 of a signature appearing in only one edition of one book and, it seems, never used as an appellation.
2. Similarly, Witztum defends an appellation used in WRR94 even though it was rejected by its own bearer, on the grounds that it is nonetheless widely used, but criticizes our use of another widely used appellation on the grounds that the bearer's son once mentioned a numerical coincidence related to a different spelling.

These are but two of many examples. Clearly, the issue of the comparative quality of the two lists, which involve historical and linguistic considerations inappropriate to this journal, cannot be broached further here. But Cohen's cited remarks, as well as work to be discussed in Section 10, support our claim to have produced a list no less rule-bound or error-free than WRR's.

Prompted by Witztum's criticisms, we adjusted our appellation list for *War and Peace* to that presented in Table 2. Compared to our original list, it is more historically accurate, performs better, and is closer to WRR's list. Note that we have removed two rabbis who have no dates in WRR's list, and one rabbi whose right to inclusion was marginal. We also added one rabbi whom WRR incorrectly excluded and imported the birth date of Rabbi Ricchi in the same way that they imported the birth date of the Besht for their first list. As in WRR94, our appellations are restricted to 5–8 letters. Detailed justifications, including responses to Witztum's critique, can be found in our updated paper (Bar-Natan and McKay, 1999) and an associated paper (Anonymous, 1999).

Several more examples of "experiments" performing well in *War and Peace* are mentioned in Section 9.

7. The study of variations

In the previous sections we discussed some of the choices that were available to WRR when they did their experiment, and showed that the freedom provided just in the selection of appellations is sufficient to explain the strong result in WRR94. Since WRR are claiming what can only be described as statistical proof of a miracle, the presence of so much "wiggle room" in the design, together with our failure to obtain any support for their claims from our own experiments (detailed in Section 10), should be sufficient reason in itself to disregard

Personality	Appellations
Rabbi Avraham Av-Beit-Din	רבי אברהם, הראב״י, הרב אב״ד, הראב״ד, האשכול
Rabbi Avraham Yitzhaki	רבי אברהם, יצחקי, זרע אברהם
Rabbi Avraham Ha-Malakh	רבי אברהם
Rabbi Aaron of Karlin	רבי אהרן
Rabbi Eliezer Ashkenazi	מעשי השם, מעשי י/ה/ו/ה, מעשי ה׳, בעל מעשי ה׳
Rabbi David Oppenheim	רבי דוד, אופנהים
Rabbi David Nieto	רבי דוד, דוד נישו
Rabbi Chaim Abulafia	רבי חיים, המהרח״א, מהרח״א
Rabbi Chaim Benbenest	רבי חיים, בנבנשתי, הרב חבי״ב, הרב החבי״ב, רב חבי״ב
Rabbi Chaim Capusi	רבי חיים, כאפוסי
Rabbi Chaim Shabtai	רבי חיים, חיים שבתי, מהרח״ש, המהרח״ש
Rabbi Yair Chaim Bacharach	חות יאיר
Rabbi Yehudah Chasid	רבי יהודה, יהודה סג״ל, הר״י חסיד
Rabbi Yehudah Ayash	רבי יהודה, מהר״י עיאש, עאיאש
Rabbi Yehosef Ha-Nagid	רבי יהוסף
Rabbi Yehoshua of Cracow	רבי יהושע, מגני שלמה
The Maharit	רבי יוסף, מטרני, יוסף טרני, טראני, מטראני, מהרימ״ט, המהרימ״ט
	מהרי״ט, המהרי״ט, הר״י טרני, הר״י טראני, ר״י טרני, ר״י טראני
Rabbi Yaacov Beirav	רבי יעקב, יעקב בירב, מהר״י בירב, הריב״ר
Rabbi Israel Yaacov Chagiz	בעל הלק״ט, מהר״י חגיז, ר״י חגיז
The Maharil	רבי יעקב, מולין, יעקב סג״ל, יעקב הלוי, מהר״י סג״ל, מהר״י הלוי, מהרי״ל, המהרי״ל
The Yaabez	היעב״ץ, הריעב״ץ, עמדין, הר״י עמדין, ר״י עמדין
Rabbi Yitzhak Ha-Levi Horowitz	רבי יצחק, הורוביץ, יצחק הלוי
Rabbi Menachem Mendel Krochmal	רבי מנחם, קרוכמאל, רבי מענדל, צמח צדק
Rabbi Moshe Zacut	רבי משה, משה זכות, מהר״ם זכות, מהרמ״ז, המהרמ״ז, המזל״ן, קול הרמ״ז
Rabbi Moshe Margalith	רבי משה, מרגלית, פני משה, מרגליות
Rabbi Azariah Figo	רבי עזריה
Rabbi Immanuel Chai Ricchi	הון עשיר, העשי״ר, אוהב ור״ע
Rabbi Shalom Sharabi	רבי שלום, שרעבי
Rabbi Shlomo of Chelm	רבי שלמה, שלמה חלמא, חעלמא
Rabbi Meir Eisenstat	רבי מאיר, איזנשטט, איזנשטאט, מהר״ם א״ש

TABLE 2. Appellations for *War and Peace*

WRR's findings. However, one can do more: there is significant circumstantial evidence that WRR's data is indeed selectively biased towards a positive result. We will present this evidence without speculating here about the nature of the process which lead to this biasing. Since we have to call this unknown process something, we will call it *tuning*.

Our method is to study variations on WRR's experiment. We consider many choices made by WRR when they did their experiment, most of them seemingly arbitrary (by which we mean that there was no clear reason under WRR's research hypothesis that they should be made in the particular way they chose to) and see how often these decisions turned out to be favourable to WRR.

Direct versus indirect tuning.

We hasten to add that we are not claiming that WRR tested all our variations and thereby tuned their experiment. This naturally raises the question of what insight we could possibly gain by testing the effect of variations which WRR did not actually try. There are two answers. First, if these variations turn out to be overwhelmingly unfavourable to WRR, in the sense that they make WRR's result weaker, the robustness of WRR's conclusions is put into question whether or not we are able to discover the mechanism by which this imbalance arose. Second, and more interestingly, the apparent tuning of one experimental parameter may in fact be a side-effect of the active tuning of another parameter or parameters.

For example, the sets of available appellations performing well for two different proximity measures A and B will not generally be the same. Suppose we adopt measure A and select only appellations optimal for that measure. It is likely that some of the appellations thus chosen will be less good for measure B, so if we now hold the appellations fixed and change the measure from A to B we can expect the result to get weaker. A suspicious observer might suggest we tuned the measure by trying both A and B and selecting measure A because it worked best, when in truth we may never have even considered measure B. The point is that a parameter of the experiment might be tuned directly, or may come to be optimized as a side-effect of the tuning of some other parameters. Fortunately for our analysis, we do not need to distinguish which possibility holds in each case. (However, we note that for the first list practically all aspects of the experiment were available for tuning, while for the second list many features had been fixed by the first list. The primary possibility for tuning of the second list was in appellation selection, but some aspects of the test method were free too.)

The space of possible variations.

Our approach will be to consider only minimal changes to the experiment. An inexact but useful model is to consider the space of variations to be a direct product $X = X_1 \times \cdots \times X_n$, where each X_i is the set of available choices for one parameter of the experiment. The model supposes that the choices could be applied in arbitrary combination, which will be close to the truth in our case. Call two elements of X *neighbours* if they differ in only one coordinate. Instead of trying to explore the whole (enormous) direct product X, we will consider only neighbours of WRR's experiment in each of the coordinate directions.

To see the value of this approach, we give a tentative analysis in the case where each parameter can only take two values. For each variation $x = (x_1, \dots, x_n) \in X$, define $f(x)$ to be a measure of the result (with a smaller value representing a stronger result). For example, $f(x)$ might be the permutation rank of P_4. A natural measure of optimality of x within X is the number $d(x)$ of neighbours y of x for which $f(y) > f(x)$. Since the parameters of the experiment have complicated interactions, it is difficult to say exactly how the values $d(x)$ are distributed across X. However, since almost all the variations we try amount to only small changes in WRR's experiment, we can expect the following property to hold almost always: if changing each of two parameters makes the result worse, changing them both together also makes the result worse. Such functions f are called *completely unimodal* (Ziegler, 1995, p. 283). In this case, it can be shown that, for the uniform distribution on X, $d(x)$ has the binomial distribution $\mathrm{Binom}(n, 1/2)$ and is thus highly concentrated near $n/2$ for large n (Williamson Hoke, 1988).

Of course, this analogy only serves as a rough guide. In reality, some of the variations involve parameters that can take multiple values or even arbitrary integer values. A few pairs of parameter values are incompatible. And so on. In addition, one can construct arguments (of mixed quality) that some of the variations are not truly "arbitrary". For these reasons, and because we cannot quantify the extent to which WRR's success measures are completely unimodal, we are not going to attempt a quantitative assessment of our evidence. We merely state our case that the evidence is strong and leave it for the reader to judge.

Regression to the mean?

"In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test – and the top group will on average fall back. This is the regression effect." (Freedman, Pisani and Purves, 1978). Variations on WRR's experiments, which constitute retest situations, are a case in point. Does this, then, mean that they should show weaker results? If one adopts WRR's null hypothesis, the answer is "yes". In that case, the very low permutation rank they observed is an extreme point in the true (uniform) distribution, and so variations should raise it more often than not. However, under WRR's (implicit) alternative hypothesis, the low permutation rank is not an outlier but a true reflection of some genuine phenomenon. In that case, there is no *a priori* reason to expect the variations to raise the permutation rank more often than it lowers it. This is especially obvious if the variation holds fixed those aspects of the experiment which are alleged to contain the phenomenon (the text of Genesis, the concept underlying the list of word pairs and the informal notion of ELS proximity). Most of our variations will indeed be of that form.

Computer programs.

A technical problem that gave us some difficulty is that WRR have been unable to provide us with their original computer programs. Neither the two programs distributed by WRR (Rosenberg, undated), nor our own independent implementations of the algorithm as described in WRR's papers (1986, 1987, 1994), consistently produce the exact distances listed in those preprints or the histograms that appear there and in WRR94. Consequently, we have taken as our baseline a program identical to the earliest program available from WRR, including its half-dozen or so programming errors. As evidence of the relevance of this program, we note that it produces the exact histograms given in WRR94 for the randomized text R, for both lists of rabbis. (The histograms for Genesis that appear in WRR94 are, according to Witztum, the results of a program, presumably lost, that preceded the one used for the permutation tests in WRR94.)

What measures should we compare?

Another technical problem concerns the comparison of two variations. Should we use the success measures employed by WRR at the time they compiled the data, or those later adopted for publication? As noted in Section 3, WRR's success measures varied over time and, until WRR94, consisted of more than one quantity. We will restrict ourselves to four success measures, chosen for their likely sensitivity to direct and indirect tuning, from the small number that WRR used in their publications.

In the case of the first list, the only overall measures of success used by WRR were P_2 and their P_1-precursor (see Section 3). The relative behaviour of P_1 on slightly different metrics depends only on a handful of $c(w, w')$ values close to 0.2, and thus only on a handful of appellations. By contrast, P_2 depends continuously on all of the $c(w, w')$ values, so it should make a more sensitive indicator of tuning. Thus, we will use P_2 for the first list.

For the second list, P_3 is ruled out for the same lack of sensitivity as P_1, leaving us to choose between P_2 and P_4. These two measures differ only in whether appellations of the form "Rabbi X" are included (P_2) or not (P_4). However, experimental parameters not subject to choice cannot be involved in tuning, and because the "Rabbi X" appellations were forced on WRR by their prior use in the first list, we can expect P_4 to be a more sensitive indicator of tuning than P_2. Thus, we will use P_4. Our choice notwithstanding, we feel that P_4 imperfectly captures WRR's probable intentions. For their experiment on the second list to have been as successful as first reported (WRR, 1986), WRR needed more than just a small value for P_2 or P_4. They also needed the distances for a cyclic shift of the dates to show a flat histogram and yield a *large* value of P_2 or P_4.

In addition to P_2 for the first list and P_4 for the second, we will show the effect of experiment variations on the least of the permutation ranks of P_{1-4}. This is not only the sole success measure presented in WRR94, but there are other good reasons. The permutation rank of P_4, for example, is a version of P_4 which has been "normalized" in a way that makes sense in the case of experimental variations that change the number of distances, or variations that tend to uniformly move distances in the same direction. For this reason, the permutation rank of P_4 should often be a more reliable indicator of tuning than P_4 itself. The permutation rank also to some extent measures P_{1-4} for both the identity permutation and one or more cyclic shifts, so it might tend to capture tuning towards the objectives mentioned in the previous paragraph. (Recall from Section 3 that WRR had been asked to investigate a "randomly chosen" cyclic shift.)

In summary, we will restrict our reporting to four quantities: the value of P_2 for the first list, the value of P_4 for the second list, and the least permutation rank of P_{1-4} for both lists. In the great majority of cases, the least rank will occur for P_2 in the first list and P_4 in the second.

The results.

Values for each of these four measures of success will be given as ratios relative to WRR's values. A value of 1.0 means "less than 5% change". Values greater than 1 mean that our variation gave a less significant result than WRR's original method gave, and values less than 1 mean that our variation gave a more significant result. Since we used the same set of 200 million random permutations in each case, the ratios should be accurate to within 10%. To save space with large numbers, we use scientific notation; for example $3e7$ means 3×10^7. The score given to each variation has the form $[p_1, r_1; p_2, r_2]$, where

$$\begin{aligned} p_1 &= \text{The value of } P_2 \text{ for the first list, divided by } 1.76 \times 10^{-9}; \\ r_1 &= \text{The least permutation rank for the first list, divided by } 4.0 \times 10^{-5}; \\ p_2 &= \text{The value of } P_4 \text{ for the second list, divided by } 7.9 \times 10^{-9}; \\ r_2 &= \text{The least permutation rank for the second list, divided by } 6.8 \times 10^{-7}. \end{aligned}$$

These four normalization constants are such that the score for the original metric of WRR is $[\mathbf{1}, \mathbf{1}; \mathbf{1}, \mathbf{1}]$. A bold "**1**" indicates that the variation does not apply to this case so there is necessarily no effect.

Two general types of variation were tried. The first type involves the many choices that exist regarding the dates and the forms in which they can be written. A much larger class of variations concerns the metric used by WRR, especially the complicated definition of the function $c(w, w')$. In both cases the details are quite technical, so we have presented them in Appendix B and Appendix C, respectively. Our selection of variations was in all cases as objective as we could manage; we did not select variations according to how they behaved.

We believe that in fact we have provided a fairly good coverage of natural minor variations to the experiment and that most qualified persons deeply familiar with the material would choose a similar set. We are happy to test any additional natural minor variation that is brought to our attention.

Conclusions.

As can be seen from the Appendices, the results are remarkably consistent: only a small fraction of variations made WRR's result stronger and then usually by only a small amount. This trend is most extreme for the permutation test in the second list, the only success measure presented in WRR94. At the very least, this trend shows WRR's result to be not robust against variations. Moreover, as explained at the beginning of this section, we believe that these observations are strong evidence for tuning, but will not attempt a quantitative evaluation.

8. Traces of naive statistical expectations

There are some cases in the history of science where the integrity of an empirical result was challenged on the grounds that it was "too good to be true" (Dorfmann, 1978; Fisher, 1965, for example); that is, that the researchers' expectations were fulfilled to an extent which is statistically improbable. Some examples of such improbabilities in the work of WRR and Gans (Gans, 1995, described in Section 9) were examined by three of the present authors (Kalai, McKay and Bar-Hillel, 1998). Here we will summarize this work briefly. It is worthy of note that these observations are surprising even if we adopt WRR's hypothesis that the codes are real.

Our interest was roused when we noticed that the P_2 value (not the permutation rank, which did not yet exist) first given by WRR for the second list of rabbis (WRR, 1987), 1.15×10^{-9}, was quite close to that of the first, 1.29×10^{-9}. To see whether this was as statistically surprising as it seemed, we conducted a Monte Carlo simulation of the sampling distribution of the ratio of two such P_2 values. This we did by randomly partitioning the total of 66 rabbis from the two lists into sets of size 34 and 32—corresponding to the size of WRR's two lists—and computing the ratio of the larger to the smaller P_2 value for each partition. Although such a random partition is likely to yield two lists that have more variance within and less variance between than in the original partition (in which the first list consisted of rabbis generally more famous than those in the second list), our simulation showed that a ratio as small as 1.12 occurred in less than one partition in a hundred. (The median ratio was about 700.)

Even under WRR's research hypothesis, which predicts that both lists will perform very well, there is no reason that they should perform equally well. This ratio is not surprising, though, if it is the result of an iterative tuning process on the second list that aims for a "significance level" (which P_2 was believed to be at that time) which matches that of the first list. Nevertheless, our observation was *a posteriori* so we are careful not to conclude too much from it.

An opportunity to further test our hypothesis was provided by another experiment that claimed to find "codes" associated with the same two lists of famous rabbis. The experiment of Gans (1995) used names of cities instead of dates, but only reported the results for both lists combined. Using Gans' own success measure (the permutation rank of P_4), but computed using WRR's method, we ran a Monte Carlo simulation as before. The two lists gave a ratio of P_4 permutation ranks as close or closer than the original partition's in less than 0.002 of all random 34-32 partitions of the 66 rabbis.

Previous research by psychologists (Tversky and Kahneman, 1971; Kahneman and Tversky, 1972) has shown that when scientists replicate an experiment, they expect the replication to resemble the original more closely than is statistically warranted, and when scientists hypothesize a certain theoretical distribution (e.g., normal, or uniform), they expect their observed data to be distributed closer to the theoretical expectation than is statistically warranted. In other words, they do not allow sufficiently for the noise introduced by sampling error, even when conditioned on a correct research hypothesis or theory. Whereas real data may confound the expectations of scientists even when their hypotheses are correct, those whose experiments are systematically biased towards their expectations are less often disappointed (Rosenthal, 1976).

In this light, other aspects of WRR's results which are statistically surprising become less so. For example, the two distributions of $c(w, w')$ values reported by WRR for their two lists (WRR, 1987; WRR94) are closer (using the Kolmogorov-Smirnoff distance measure) than 97% of distance distributions, in a Monte Carlo simulation as before.

As a final example, when testing the rabbis lists on texts other than Genesis, WRR were hoping for the distances to display a flat histogram. Some of the histograms of distances they presented (WRR, 1987) were not only gratifyingly flat, they were surprisingly flat: two out of the three histograms presented in that preprint are flatter than at least 98% of genuine samples of the same size from the uniform distribution. A similar story can be told about the distances for the cyclic shift of the dates (see Section 3). The details can be found in Kalai, McKay and Bar-Hillel (1998).

It is clear that some of these coincidences might have happened by chance, as their individual probabilities are not extremely small. However, it is much less likely that chance explains the appearance of all of them at once. As a whole, the findings described in this section are surprising even under WRR's research hypothesis and give support to the theory that WRR's experiments were tuned towards an overly idealized result consistent with the common expectations of statistically naive researchers.

9. Additional claims of ELS phenomena

The truth about controversial claims in science is seldom resolved merely by close inspection of the experiments which lead to them. Much more important is whether the phenomenon persists under replication. In this section we discuss other experiments that claim to provide support for WRR's theory. In the next section, we describe some of the many experiments we have performed ourselves.

Two further (unpublished) papers of WRR exist (WRR, 1995; WRR, 1996) describing seven experiments altogether and reporting a positive result for all but one of them. The single negative result (an experiment on female names which mimiced another on male names) is the only experiment reported by WRR which had no freedom of movement in its design.

The 70 nations experiment.

Even by their own account, the most impressive experiment of WRR other than that in WRR94 is the "70 nations" experiment (WRR, 1995), which concerns the list of nations of the world in Chapter 10 of Genesis. The word pairs are of two types.

One type of word pair consists of the name of a nation and that same name with one of four attributes attached. For example, there is the pair ⟨"Gomer," "language of Gomer"⟩. The four attributes used are alleged to have been derived *a priori* from the writings of the Vilna Gaon (a great rabbi of the eighteenth century). However, two of the four attributes do not appear there. Instead, the Vilna Gaon uses other words, including a different Hebrew word for "language of", that do not perform well at all.

To illustrate the great freedom available in producing "experiments" of this nature, Bar-Natan, McKay and Sternberg (1998) present a different set of four attributes, found in the writings of the Ramban (Nachmanides, a great rabbi of the thirteenth century), which produce a result in *War and Peace* 100 times better than WRR's attributes produce in Genesis (using the same method of analysis). This time, however, all four attributes appear in the source. Several other examples that illustrate the same point are presented by Bar-Natan, McKay and Sternberg as well, even a natural set of five attributes that gives a strong "significance level" in *both* Genesis and *War and Peace*.

The other type of word pair in the 70 nations experiment consists of the name of a nation and another associated word. We will not go into the details here; suffice it to say that the associated words were chosen in an unsystematic manner from a larger set, with very many arbitrary decisions invariably made in the favourable direction.

Header samples.

Another class of experiment presented by WRR is the "header sample," where one word is matched against a small collection of related words. A number of examples appear in a preprint of WRR (1996) and are characterized by inconsistent application of ad hoc rules. A more recent example (relying on an invalid method of randomization) appears in an Internet article of Witztum (1998b). We have found that constructing convincing examples of this type of "experiment" in any text is easy (Em Piqchit, 1998; McKay *et al.*, 1998), and see no reason to take them seriously. A discussion of how such experiments can be constructed is given by McKay (1998).

The cities experiment of Gans.

The only other significant claim for a positive result is the preprint of Gans (1995), which analyzes data given to him by an associate of Witztum. Gans uses the names of the cities of birth or death of the famous rabbis in place of their dates. It was later withdrawn (Gans, 1998), but Gans recently announced a new edition which we have not yet seen. The original edition raises our concerns regarding the objectivity of the cities data, as many choices were available. We also note that the variations of the metric that we describe in Appendix C cause deterioration of Gans' result (1995) just as regularly as for WRR's experiment. Also see the following section for a similar experiment of ours that failed to find any phenomenon.

10. Independent ELS experiments

We have, of course, conducted many real experiments of our own on the Bible. These were sincere attempts at replication and are not to be confused with the demonstrations of data manipulation we mentioned in Sections 6 and 9. In designing our experiments, we strove for specifications which were as simple and complete as possible, allowing a bare minimum of "wiggle room" in the collection of the data. In some cases it was impossible to avoid an amount of arbitrary, though *a priori*, choice.

Despite our concerns with WRR's experimental method, we felt obligated to use it ourselves, even in our own experiments. The reason is simply that this is the method by which WRR claim "codes" to be detectable. Our failure to detect them by the same method is a negative result directly bearing on WRR's claims without regard to the problems that the method has. However, since WRR's null hypothesis is not true conditional on the list of word pairs, we must always bear in mind the possibility that to some extent we are measuring some subtle mundane correlations in the data. For example, in the cities experiment mentioned in the previous section, how do we account for the obvious correlation between names and places of birth? WRR and Gans are silent on these issues.

Our own rabbis experiments.

Perhaps the most important class of experiments we have conducted are repetitions of the famous rabbis experiment. For this purpose, we engaged Simcha Emanuel, a specialist in rabbinical history at Tel-Aviv University, as an independent consultant.

For the first experiment, Emanuel was informed which 32 rabbis appeared on WRR's second list and asked to prepare names and appellations for each of them. He had not seen WRR's lists and was asked not to consult them, nor was he given any explicit guidance concerning which types of appellations to include and how to spell them. Rather, he was asked to use his own professional judgement to settle all issues. During his work he consulted a second historian, David Assaf of Tel-Aviv University. As well as writing names and appellations, Emanuel and Assaf commented on the accuracy of the dates given by Margaliot (1962) and corrected some of them (as had WRR).

The result of this experiment was a list of names and appellations which appears quite different from that of WRR. The least permutation rank of P_{1-4} was 0.233.

The same exercise was then carried out with a list of rabbis that had not been used before, namely those whose entries in Margaliot's encyclopedia occupy from 1 to 1.5 columns and for whom there is a date of birth or death mentioned (except for those incorrectly included by WRR in their second list). For these 26 rabbis, the least permutation rank of P_{1-4} was 0.404.

After the above two experiments were completed, we carried out the following re-enactment of WRR's second experiment.

1. A list of rabbis was drawn from Margaliot's encyclopedia by applying WRR's criteria for their second list, while correcting the errors they made. Our list differed from WRR's in dropping two rabbis and including three others. One rabbi who fits the selection criteria could not be included because he appears incorrectly in WRR's first list.
2. Emanuel was shown the spelling rules and table of appellations for WRR's first list as they first appeared in WRR (1986). He then compiled a parallel table of appellations for our list of 33 rabbis, attempting to follow the rules and practices of WRR's first list.
3. To mimic WRR's processing of dates for their first list, we used the dates given by Margaliot except in the cases where Emanuel either found an error or found an additional date. In some cases Emanuel regarded a date as uncertain, in which case we followed WRR's practice of leaving the date out. Overall, Emanuel changed more of Margaliot's dates than WRR did.
4. The resulting list of word pairs was processed using WRR's permutation test.

The result of applying WRR's permutation test was that the least permutation rank of P_{1-4} was an uninteresting 0.254.

There are some syntactic differences between Emanuel's list and WRR's first list, namely that Emanuel was sparing in use of articles and sometimes used a one-letter abbreviation for "Rabbi". We pointed out these differences to Emanuel, who then made some changes to his list. Because of our intervention, the new list cannot be said to be as *a priori* as the original, but it is arguably closer to the practices of WRR's first list. The new list gives permutation ranks of 0.154, 0.054, 0.089, and 0.017 for P_{1-4}, respectively. Applying the Bonferroni inequality as in WRR94, we have an overall significance level of 0.066.

This negative result is all the more conclusive if we realize that our experiment had some clear biases towards WRR's experiment. The definition of the set of rabbis, the introduction of P_3 and P_4 (only P_1 and P_2 appeared with the first list) and, most importantly, the definition of the permutation test, were under WRR's control when they ran their second experiment and were merely copied by us. Thus, we were vulnerable to any systematic bias that existed in those decisions, as well as to the possibility that WRR knew some examples from their second list earlier than acknowledged. We can only partly compensate for these biases. Using only P_1 and P_2 changes the overall result to 0.108. Using the permutation test of Diaconis (discussed in Sections 3 and 4) rather than the test invented by WRR, the results are even worse: 0.647 using the average and 0.743 using the minimum.

We believe that these experiments clearly establish that the success of WRR's experiment was primarily due to the choices made in compiling their lists and not to any genuine ELS phenomenon in Genesis. The data for the above three experiments can be found at McKay's web site (1999b).

Replication of Gans' experiment.

The experiment of Gans (1995), which used the cities of birth and death of the famous rabbis in place of the dates, prompted Barry Simon of Caltech to design the following more objective variant (Simon, 1998): use the names of all the cities mentioned in each rabbi's entry in Margaliot's encyclopedia as places of birth, death, living, working or studying, without any modification of spelling or addition of prefixes. This data was matched against WRR's appellations. The least permutation rank out of P_{1-4} was 0.133 for the first list and 0.324 for the second. The same experiment using *Encyclopedia Hebraica* (1988) in place of Margaliot (these were the two sources used by Gans, 1995) produced 0.324 and 0.052, respectively. Then we adopted the following procedures of Gans: use only cities of birth and death, combine the two lists, use only P_4 and allow two prefixes meaning "community of". This version also failed: 0.550 for Margaliot's encyclopedia and 0.117 for the Hebraica.

It is difficult to escape the conclusion that the result of Gans (1995) also reflects more on the data than on any phenomenon inherent in Genesis.

Other replications.

As mentioned in Section 5, WRR used only some of the possible ways to write dates. The additional ways can be considered to be independent replications which are exceptionally tight in their design. The results were all negative, as we showed in Table 1.

Another obvious example of a replication was inspired by the fact that WRR only used the day and month of birth and death, not the year. Bar-Natan, Gindis and McKay (1999) performed an experiment using the year instead, as well as one using the names of famous books written by each rabbi, in place of his dates. The lists of years and books were extracted from the above-mentioned two encyclopedias according to simple mechanical rules publicized in advance. For the years of birth and death, there were three ways of writing the years and two ways of analysing them (permutation ranks of P_1 and P_2). Thus, there were six "significance levels" for each list of rabbis, the smallest being 0.050 and 0.053, respectively. For the books, there were two "significance levels" for each list, of which the smallest were 0.981 and 0.228, respectively.

In another experiment, we verified that the other four books of the Torah (Pentateuch), relative to which Genesis holds no special privilege in Jewish tradition, show no "encoding" of WRR's lists of rabbis. The details appear in Table 3. Note that there is a dependency between these permutation ranks and the permutation ranks for Genesis, due to the fact that the distribution of permutation ranks conditioned on the word list is not uniform (see Section 4). It may be that a low permutation rank in Genesis enhances the probability of a low permutation rank in other books.

Book	List 1	List 2
Exodus	0.0212	0.1010
Leviticus	0.6950	0.8467
Numbers	0.0046	0.6628
Deuteronomy	0.0664	0.7282

TABLE 3. Other books of the Torah. (Least of P_{1-4} permutation ranks)

Another experiment was suggested by Rips' observation that the name of Theodor Herzl (a famous Zionist) appears close to his birth date. The names (family names and full names) of all presidents, prime ministers, and Knesset speakers of the State of Israel from 1948 to the present were matched against the dates of their birth, their first inauguration into office, and (if known) their death. A complete definition of the data was made in advance by James Price of Temple Baptist Seminary. Dates (day, month, and year) were written according to instructions provided by the Academy for the Hebrew Language. The resulting data (see McKay, 1999b) gave a permutation rank of 0.512 for P_1 and 0.768 for P_2. Using only the day and month written using WRR's rules, the permutation ranks were 0.155 for P_1 and 0.044 for P_2. In other words, the experiment was another failure.

In addition, we have performed many other ELS experiments, all of which failed to detect anything unusual. Here we mention a few more examples.

1. The experiment of Gans (1995) excluded the appellations with the form "Rabbi X". We tried it with *only* those appellations. The least permutation rank was 0.079.
2. An experiment was constructed to test whether the various appellations for the same rabbi tend to appear close to each other (cf. Figure 1 in WRR94). In order to be able to apply WRR's method, we randomly divided the set of appellations for each rabbi into two subsets, and measured the distances between the appellations in one subset and the appellations in the other. Ten such random partitions were tried, of which the smallest permutation rank observed was 0.179 for the first list and 0.129 for the second list. We also tried simply replacing each set of dates by duplicates of the set of appellations, with $c(w, w')$ being treated as undefined if $w = w'$. The result was a least permutation rank of 0.872 for the first list and 0.573 for the second list.
3. As part of our investigation of the 70 nations experiment (see Section 9), we made an *a priori* list of 132 additional attributes of nations, and tried them all in Genesis. Both the overall distribution of the 132 permutation ranks, and the magnitude of the extreme values, were consistent with what we observed for *War and Peace* and for randomized texts.

In summary, despite a considerable amount of effort, we have been unable to detect the "codes". This is in stark contrast to the near-perfect reported success rate of WRR.

11. The matter of the text

Popular accounts of the "Bible code" almost inevitably speak of ELS-encoded information in the "original text" of the Bible, often with a claim that the "original text" is the one used in WRR94. However, WRR94 used an edition based on several differing sources, that was published in 1962 by Koren Publishers. Due to the importance of this issue, we will briefly summarize what is known about the history of the text and what the consequences of this history are.

One of the characteristics of written Hebrew is its inconsistency in the use of vowel letters (known as *matres lectionis*). Words can be spelt without vowel letters ("defective spelling"), with them ("full spelling"), or in some mixture thereof. The earliest known Hebrew inscriptions, dating from the tenth century BC, use defective spelling almost exclusively. However, all versions of all books of the Hebrew Bible known to us today employ a complex mixture of full and defective spelling, not even consistently for the same words. The Babylonian Talmud (Kiddushin 30a), written around the fourth century AD, reports that full knowledge of the original spellings had by that time been lost. This evidence, and much other evidence, has led most scholars to believe that either one or more major revisions, or a long gradual process of slow revision, produced major changes in the letter-by-letter text between the original and the first historic editions (Cross and Freedman, 1952; Naveh, 1987; Zevit, 1980).

The Dead Sea Scrolls date from the third century BC to the first century AD. There are many scrolls of Genesis amongst them, but they have survived only in small fragments. The rate of variation between the surviving fragments and the present text varies from about 1 letter in 1200 to about 1 letter in 20 (Tov, 1998; Ulrich *et al.*, 1994). Because of the wide range of textual variants, and for other reasons, the general consensus amongst experts is that the scrolls are representative of the textual situation throughout Palestine at the time (Cohen, 1998; Tov, 1992). The amount of variation that had already occurred during the many preceding centuries since Genesis was written is a matter of scholarly speculation, though the considerations of the previous paragraph suggest it was very great.

Around the eight to tenth centuries, there was a major process of standardization leading to the so-called Masoretic text, which displaced most other extant editions within the next few hundred years. Still, despite the exercise of very great care, the difficulty of exact copying by hand is so great that we do not have two identical scrolls from before the advent of printed editions in the sixteenth century. Differences between scrolls could amount to anything from a few letters to thousands of letters.

For extensive information on this subject, see Breuer (1976), Cohen (1998), Tigay (1998) and Tov (1998).

In summary, there is hardly any chance that the Koren edition is close in letter-by-letter detail to the original text. In fact, if the text of Genesis were to be consistently spelled in the style of the inscriptions dated closest to the traditional year when Genesis was written, the differences would number in the thousands (even without any change of meaning). This conclusion has catastrophic consequences for any theory that "codes" in the original text have survived until today. Clearly an ELS is destroyed if any letter is inserted or deleted within its overall span. The ELSs giving the strongest contribution to the WRR94 result together span most of the text. Our experiments show that deletion of 10 letters in random places is enough to degrade the result by an average factor of 4000, and deleting 50 letters is enough to eliminate it completely. Of course, the effect has a very large variance, as it depends on which of a comparatively small number of important ELSs are "hit" by a deletion. The first list is even more sensitive to the effects of such corruption, as its important ELSs have greater skip. Ten letters deleted in random places are on average enough to eliminate its significance altogether.

To further explore the effect of textual corruption, we performed WRR's experiment on a number of other exemplary Genesis texts, prepared for us by Jeffrey Tigay (Professor of Hebrew and Semitic Languages and Literatures at the University of Pennsylvania) using information provided by Menachem Cohen. These include the Yemenite edition, which is probably the best single representative we have today of the Masoretic text of Genesis, and the Leningrad Codex, which is the oldest complete text of the Hebrew Bible still in existence. For each text, Table 4 gives the number of letters by which it differs from the Koren edition (see Cohen, 1997a, for details) and the best permutation rank out of P_{1-4}. Because it is believed by experts (Breuer, 1976; Cohen, 1997b) that the Yemenite edition is more likely to be correct than the Koren edition in each of the three places where they differ, we also show the effect of each of those differences separately. The rows labelled Koren-1 to Koren-3 show the Koren text with each of the three single changes applied.

Editions	Differences	List 1	List 2
Koren (WRR)	0	0.000038	0.0000006
Koren-1	1	0.000317	0.0000022
Koren-2	1	0.000106	0.0000008
Koren-3	1	0.000146	0.0000030
Yemenite	3	0.001421	0.0000019
Sassoon	11	0.428413	0.000231
Venice Mikraot Gedolot	15	0.029184	0.001661
Leningrad Codex	22	0.007574	0.001253
Jerusalem	35	0.008234	0.001907
Hilleli	43	0.002124	0.000641

TABLE 4. Other editions of Genesis. (Least of P_{1-4} permutation ranks)

We can see that the Koren edition, the one used by WRR, is a clear winner for both lists. It is even true that each of the probable three errors in the Koren edition contribute to WRR's advantage in both lists.

It may be noticed that the values in Table 4 appear to contradict the experiments we did on random corruption. The explanation is partly that the variance of the effect is large (as the table shows clearly), and partly that our experiments used random deletions (because that is the reverse of the general historic trend). The differences counted in the table comprise a mixture of deletions and insertions, and these tend to cancel each other out somewhat.

One rather ingenious argument that has been advanced to handle the problem of textual corruption runs as follows: the "codes" that we see today are merely a remnant of a much more perfect phenomenon that existed in the original text. The major problem with this argument is that the total amount of divergence from the original text has probably been enough to obliterate any perfect pattern several times over, not merely to dilute it. Moreover, there is a way to test the argument experimentally. If textual corruption occurred more or less in random places, it would have preferentially destroyed large-span ELSs more than short-span ELSs. However, this does not match the evidence. This is especially clear for the first list of rabbis: if the experiment of WRR is preserved in every way except that ELSs spanning more than 2000 letters (from the first to the last letters of the ELS) are ignored, the permutation ranks of P_{1-4} are all greater than 0.15. In other words, the "phenomenon" is based in large part on ELSs that present easy targets to the process of textual corruption. Similarly, for the same list without a span limit, there is no detectable correlation between $c(w, w')$ and $\max(s(w), s(w'))$, where $s(w)$ is the least span of an ELS of w. The Spearman rank correlation statistic is only 0.015. Thus, there is good evidence against the conjecture that the present "codes" are the remnants of earlier more perfect "codes".

12. Conclusions

In the words of editor Robert Kass, the paper of WRR was presented in this journal as a "challenging puzzle". The single most baffling part of the puzzle was the fact that WRR, "[i]n order to avoid any conceivable appearance of having fitted the tests to the data," produced a "fresh sample, without changing anything else" (WRR94), but nevertheless obtained a remarkable result.

The solution to the puzzle lies in considering, not fitting of the tests to the data, but fitting of the data to the tests. Not only did we identify the unacknowledged source of the flexibility (primarily the fact that the available set of appellations for the famous rabbis is more than twice as large as the set actually used), but we proved that this flexibility is enough to allow a similar result in a secular text. We supported this claim by observing that, when the many arbitrary parameters of WRR's experiment are varied, the result is usually weakened, and also by demonstrating traces of naive statistical expectations in WRR's experiment.

Be that as it may, our most telling evidence against the “codes” is that we cannot find them. All of our many earnest experiments produced results in line with random chance. These included a re-enactment of the famous rabbis experiment with the help of independent experts, Emanuel and Assaf.

In light of these findings, we believe that Kass’ “challenging puzzle” has been solved.

Appendix A. The metric defined by WRR

In order to understand some of the technical parts of this paper, especially Appendix C, it is necessary to know the details of WRR's method of calculating distances. In this Appendix we give a concise definition of the metric $c(w, w')$. We will always consider a fixed text $G = g_1 g_2 \cdots g_L$ of length L.

WRR's basic method for assessing how a word appears with equal spacing in the text (i.e., as an ELS) is to seek it also with slightly unequal spacing. These *perturbed ELSs* have all their spacings equal except that the last three spacings may be larger or smaller by up to 2. Formally, consider a word $w = w_1 w_2 \cdots w_k$ of length $k \geq 5$ and a triple of integers (x, y, z) such that $-2 \leq x, y, z \leq 2$. An (x, y, z)*-perturbed ELS* of w, or (x, y, z)*-ELS* for short, is a triple (n, d, k) such that $g_{n+(i-1)d} = w_i$ for $1 \leq i \leq k-3$, $g_{n+(k-3)d+x} = w_{k-2}$, $g_{n+(k-2)d+x+y} = w_{k-1}$ and $g_{n+(k-1)d+x+y+z} = w_k$.

It is seen that a $(0, 0, 0)$-ELS is merely a substring of equally spaced letters in the text that form w; that is, an ELS as we previously defined it. Other values of (x, y, z) represent nonzero *perturbations* of the last three letters from their natural positions. Including $(0, 0, 0)$, there are 125 such perturbations.

In measuring the properties of an (x, y, z)-ELS, there is a choice of using the perturbed or unperturbed letter positions. For example, the last letter has perturbed position $n + (k-1)d + x + y + z$ and unperturbed position $n + (k-1)d$. The paper WRR94 is unclear on this point, but we know from WRR's programs (Rosenberg, undated) that the unperturbed positions were used. Thus, we require that $g_{n+(k-1)d+x+y+z} = w_k$, according to the definition of (x, y, z)-ELS, but when we measure distances we assume the letter is really in position $n + (k-1)d$.

To continue to the next step, we define the *cylindrical distance* $\Delta(t, h)$. Roughly speaking, it is the shortest distance, along the surface of a cylinder of circumference h, between two letters that are t positions apart in the text, when the text is written around the cylinder. However, this is only approximately correct. The definition of $\Delta(t, h)$ given in WRR94 is not exactly what they used, so we give the definition WRR gave earlier (1986) and in their programs (Rosenberg, undated). Define the integers Δ_1 and Δ_2 to be the quotient and remainder, respectively, when t is divided by h. (Thus, $t = \Delta_1 h + \Delta_2$ and $0 \leq \Delta_2 \leq h-1$.) Then

$$\Delta(t,h)^2 = \begin{cases} \Delta_1^2 + \Delta_2^2, & \text{if } 2\Delta_1 \leq h; \\ (\Delta_1 + 1)^2 + (\Delta_2 - h)^2, & \text{otherwise.} \end{cases}$$

Now consider two (x,y,z)-ELSs, $e = (n,d,k)$ and $e' = (n',d',k')$. For any particular cylinder circumference h, define

$$\begin{aligned}\delta_h(e,e') &= \Delta(d,h)^2 + \Delta(d',h)^2 + \min_{0\le i\le k-1,\, 0\le i'\le k'-1} \Delta(|n+di-n'-d'i'|,h)^2 \\ \mu_h(e,e') &= 1/\delta_h(e,e').\end{aligned}$$

The third term of the definition of $\delta_h(e,e')$ is the closest approach of a letter of e to a letter of e'.

The next step is to define a multiset $H(d,d')$ of values of h. For $1 \le i \le 10$, the nearest integers to d/i and d'/i ($\frac{1}{2}$ rounded upwards) are in $H(d,d')$ if they are at least 2. Note that $H(d,d')$ is a multiset; some of its elements may be equal. Given $H(d,d')$, we define

$$\sigma(e,e') = \sum_{h\in H(d,d')} \mu_h(e,e').$$

For any (x,y,z)-ELS e, consider the intervals I of the text with this property: I contains e, but does not contain any other (x,y,z)-ELS of w with a skip smaller than d in absolute value. If any such I exists, there is a unique longest I; denote it by T_e. If there is no such I, define $T_e = \emptyset$. In either case, T_e is called the *domain of minimality* of e. Similarly, we can define $T_{e'}$. The intersection $T_e \cap T_{e'}$ is the *domain of simultaneous minimality* of e and e'. Define $\omega(e,e') = |T_e \cap T_{e'}|/L$.

Next define a set $E^{(x,y,z)}(w)$ of (x,y,z)-ELSs of w. Let D be the least integer such that the expected number of ELSs of w with absolute skip distance in $[2,D]$ is at least 10, for a random text with letter probabilities equal to the relative letter frequencies in G, or ∞ if there is no such integer. Then $E(w) = E^{(x,y,z)}(w)$ contains all those (x,y,z)-ELSs of w with absolute skip distance in $[2,D]$. Note that the formula $(D-1)(2L-(k-1)(D+2))$ in WRR94 for the number of potential ELSs for that range of skips is correct, but WRR's programs (Rosenberg, undated) use $(D-1)(2L-(k-1)D)$. We will do the same. Next define

$$\Omega^{(x,y,z)}(w,w') = \sum_{e\in E(w),\, e'\in E(w')} \omega(e,e')\sigma(e,e'),$$

provided $E(w)$ and $E(w')$ are both non-empty. If either is empty, $\Omega^{(x,y,z)}(w,w')$ is undefined.

Now, finally, we can define $c(w,w')$. If there are less than 10 values of (x,y,z) for which $\Omega^{(x,y,z)}(w,w')$ is defined, or if $\Omega^{(0,0,0)}(w,w')$ is undefined, then $c(w,w')$ is undefined. Otherwise, $c(w,w')$ is the fraction of the defined values $\Omega^{(x,y,z)}(w,w')$ that are greater than or equal to $\Omega^{(0,0,0)}(w,w')$.

In summary, by a tortuous process involving many arbitrary decisions, a function $c(w, w')$ was defined for any two words w and w'. Its value may be either undefined or a fraction between 1/125 and 1. A small value is regarded as indicating that w and w' are "close".

Appendix B. Variations of the dates and date forms

This Appendix gives the technical details for the first collection of variations we tried on the experiment of WRR, namely those involving the dates and the ways that dates can be written.

We begin with some choices directly concerning the date selection. WRR had the option of ignoring the obsolete ways of writing 15 and 16. This variation gets a score of $[8.7, 2.7; 33, 5.2]$ (in other words, omitting those forms would have made the four measures weaker by those factors). They could have written the name of the month Cheshvan in its full form Marcheshvan, $[6.4, 1.8; 96, 51]$, or used both forms, $[1.0, 1.0; 1.0, 1.0]$. They could have spelt the month Iyyar with two *yod*s on the basis of a firm rabbinical opinion (*Encyclopedia Talmudica*, 1992), $[7.2, 1.9; 3.7, 4.0]$, or used both spellings, $[\underline{0.3}, 1.1; 5.5, 5.6]$. (We will underline all values less than 1 to help the reader appreciate how few they are.) They could have written the two leap-year months Adar 1 and Adar 2 as Adar First and Adar Second instead (as their source, Margaliot, usually does), $[9.2, 6.1; 1.0, 1.0]$, or used both forms, $[\underline{0.8}, \underline{0.9}; 1.0, 1.0]$. Note that each of these variations only applies to a few rabbis.

A more drastic variation available to WRR was to use the names of months that appear in the Bible, which are sometimes different from the names used now. Those names are: Ethanim, Bul, Kislev, Tevet, Shevat, Adar, Nisan, Aviv (another name for Nisan), Ziv, Sivan, Tammuz and Elul. The month of Av is not named at all. This variation gives a score of $[220, 24; 3400, 2800]$ if the Biblical names are used alone (with two names for Nisan and none for Av) and $[1.7, 10.5; 67, 450]$ if both types of name are used together. This variation is consistent with WRR's frequently stated preference for Biblical constructions.

As an aside, a universal truth in our investigation is that whenever we use data completely disjoint from WRR's data the phenomenon disappears completely. For example, we ran the experiment using only month names (including the Biblical ones) that were *not* used by WRR, and found that none of the permutation ranks were less than 0.11 for any of P_{1-4}, for either list.

WRR were inconsistent in that for their first list they introduced a date not given (even incorrectly) by Margaliot, whereas for their second list they did not. They could have acted for the first list as they did for the second (i.e., not introduce the birth date of the Besht), $[8.2, 4.9; \mathbf{1}, \mathbf{1}]$. Alternatively, they could have imported other available dates into the second list. For example, Rabbi Emdin was born on 15 Sivan (Bik, 1974; Schacter, 1988), $[\mathbf{1},\mathbf{1};\underline{0.3},\underline{0.3}]$, Rabbi Ricchi on 15 Tammuz (Vilenski, 1949), $[\mathbf{1},\mathbf{1};\underline{0.3},2.6]$, and Rabbi Yehosef Ha-Nagid on 11 Tishri (Ha-Nagid, 1926), $[\mathbf{1},\mathbf{1};1.0,3.9]$. They could have used the doubt about the death date of Rabbenu Tam (discussed at length by Reiner, 1997, p. 7) to remove it, as they did with other disputed dates, $[1.6,\underline{0.7};\mathbf{1},\mathbf{1}]$, or similarly for Rabbi Chasid (Gedaliah, 1963), $[\mathbf{1},\mathbf{1};1.0,1.5]$. They could have used the correct death date of Rabbi Beirav (1 Iyyar; see Mabit), $[\mathbf{1},\mathbf{1};1.3,\underline{0.8}]$ or the correct death date of Rabbi Teomim (10 Iyyar; Teomim, 1993), $[\mathbf{1},\mathbf{1};\underline{0.9},1.2]$.

They could also have written all the dates in alternative valid ways. The most obvious variation would have been to add the form akin to "on 1st of May". It gives the score $[1.2,2.2;\underline{0.6},16.4]$.

The eight regular date forms in Table 1 can be used in $2^8-1=255$ non-empty combinations of which WRR used one combination (i.e., the first three). We tried all 255 combinations, and found that WRR's choice was uniquely the best for the first and fourth of our four success measures. In the case of our second measure (least permutation rank of P_{1-4} for the first list), WRR's choice is sixth best. (The best is a subset of their three forms.) For our third measure (P_4 for the second list), WRR's choice is third best. Since the various date forms are not equal in their frequency of use, it would be unwise to form a quantitative conclusion from these observations.

Appendix C. Variations of the metric

This Appendix gives the technical details for the variations we tried on WRR's method of analysis. In all cases presented here, the text of Genesis and the list of word pairs was held fixed. A deep understanding of the metric is needed for this Appendix, for which we refer the reader to Appendix A.

First consider the function $\delta_h(e,e')$ that lies at the heart of the WRR metric. Define these quantities:

$$
\begin{aligned}
f &= \Delta(d,h),\\
f' &= \Delta(d',h),\\
l &= \min \Delta(|n+di-n'-d'i'|,h),\\
\mu &= \operatorname{mean} \Delta(|n+di-n'-d'i'|,h),\\
m &= \Delta(|2n+d(k-1)-2n'-d'(k'-1)|/2,h),\\
L &= \max \Delta(|n+di-n'-d'i'|,h),\\
x,y &= \text{dimensions of smallest enclosing rectangle},
\end{aligned}
$$

where the min, mean, and max are taken over $0 \leq i \leq k-1$ and $0 \leq i' \leq k'-1$. The quantity m is the cylindrical distance between the midpoints of the two ELSs.

$\phi(e, e')$	$\delta(e, e') = \phi(e, e')$	$\delta(e, e') = \sqrt{\phi(e, e')}$
$f^2 + f'^2 + l^2$	[**1**, **1**; **1**, **1**] (WRR)	[154, 120; 10.1, 99]
$f^2 + f'^2 + m^2$	[1.5, 3.7; 66, 92]	[65, 83; 101, 650]
$f^2 + f'^2 + \mu^2$	[1.3, 5.1; $\underline{0.6}$, 2.3]	[168, 230; 25, 410]
$f^2 + f'^2 + L^2$	[2.4, 4.1; 1.0, 11.4]	[220, 340; 40, 1000]
$f^2 + f'^2 + 2l^2$	[2.5, 1.6; 2.8, 1.1]	[210, 88; 12.1, 66]
$2f^2 + 2f'^2 + l^2$	[1.4, 1.3; $\underline{0.6}$, 1.8]	[61, 82; 11.7, 220]
$(f + f' + l)^2$	[1.8, 1.9; $\underline{0.5}$, 1.0]	[190, 137; 10.1, 154]
$(f + f' + m)^2$	[$\underline{0.6}$, 1.9; 17.5, 57]	[98, 120; 130, 1200]
$(f + f' + \mu)^2$	[3.6, 8.3; $\underline{0.4}$, 3.7]	[220, 290; 20, 550]
$(f + f' + L)^2$	[7.1, 15.1; $\underline{0.5}$, 11.6]	[430, 460; 34, 1100]
$\max(f, f', l)^2$	[2.4, 1.3; 2.7, 1.9]	[86, 76; 6.8, 69]
$\max(f, f', m)^2$	[3.9, 6.8; 240, 230]	[40, 58; 74, 400]
$\max(f, f', \mu)^2$	[2.9, 9.8; 1.2, 3.0]	[220, 280; 25, 310]
$\max(f, f', L)^2$	[2.5, 13.3; 1.1, 12.1]	[380, 500; 39, 810]
μ^2	[5.7, 18.6; 2.2, 4.2]	[340, 360; 49, 420]
L^2	[2.8, 13.6; 1.3, 12.3]	[420, 530; 35, 740]
$(L + l)^2$	[4.0, 13.8; 2.1, 7.0]	[360, 380; 73, 570]
$L^2 + l^2$	[2.7, 13.4; $\underline{0.9}$, 5.5]	[330, 450; 38, 600]
$(x + y)^2$	[30, 44; $\underline{0.5}$, 16.8]	[640, 550; 15.5, 630]
$x^2 + y^2$	[15.1, 33; $\underline{0.4}$, 9.7]	[500, 610; 18.5, 620]
$\max(x, y)^2$	[9.9, 31; $\underline{0.2}$, 5.9]	[190, 340; 31, 840]
xy	[680, 140; $\underline{0.5}$, 71]	[1.1e4, 720; 97, 3900]
$x^2 + y^2 + l^2$	[8.9, 26; $\underline{0.4}$, 4.7]	[180, 320; 24, 740]
$x^2 + y^2 + m^2$	[1.5, 13.2; 2.3, 14.4]	[150, 340; 26, 830]
$x^2 + y^2 + \mu^2$	[7.4, 24; $\underline{0.5}$, 5.4]	[183, 310; 23, 680]
$x^2 + y^2 + L^2$	[14.7, 38; $\underline{0.7}$, 8.2]	[430, 560; 27, 720]
$(x + y + l)^2$	[7.1, 17.4; $\underline{0.1}$, 1.1]	[250, 290; 21, 440]
$(x + y + m)^2$	[2.0, 13.7; 1.9, 13.5]	[230, 380; 28, 705]
$(x + y + \mu)^2$	[22, 22; $\underline{0.3}$, 4.3]	[430, 500; 22, 650]
$(x + y + L)^2$	[10.4, 26; $\underline{0.8}$, 12.6]	[610, 630; 37, 1100]
$xy + l^2$	[42, 28; $\underline{0.3}$, 1.4]	[3900, 600; 46, 211]
$xy + m^2$	[4.0, 17.3; 3.8, 26]	[670, 440; 74, 830]
$xy + \mu^2$	[11.6, 27; $\underline{0.4}$, 3.2]	[740, 560; 49, 650]
$xy + L^2$	[9.4, 26; $\underline{0.9}$, 15.0]	[810, 710; 43, 1050]

TABLE 5. The effect of changing $\delta_h(e, e')$

WRR define $\delta_h(e,e') = f^2 + f'^2 + l^2$, which is a square of a distance. In Table 5 we show the effects of making other choices. We have restricted ourselves to distances and squares of distances, and to functions which measure the same type of compactness that WRR's function measures. The latter condition is enforced in a strong sense: for bounded word length, each function in Table 5 is bounded above and below by moderate constant multiples of the first. For example, $f^2 + f'^2 + l^2 \le (f + f' + l)^2 \le 3(f^2 + f'^2 + l^2)$.

The paucity of values less than 1 in the table and their blandness is remarkable. We did not find a single variation that improved the result of the permutation test for either list. In the case of the first list, only one variation improved P_2, and then only by a little. Only the P_4 value for the second list shows a significant number of improvements (19 out of 67), which is not too surprising in light of the fact that P_4 was not the only criterion of success. In this regard, we mention that only 6 of the 67 variations in the table increase the value of P_4 for the distances after the cyclic shift of the dates (another of WRR's success measures, but one they wanted to be large; see Sections 3 and 7). Similarly, only 4 of the 67 variations improve the flatness of the histogram of those distances, as measured by the χ^2 statistic with 25 bins (the same bins displayed in WRR94).

Furthermore, in all 19 cases where P_4 dropped, the permutation rank of P_4 increased. This indicates that the observed drop in P_4 values is due to an overall tendency for $c(w,w')$ values to decrease when these variations are applied. In other words, it is an example of the inadequacy of P_4 as an indirect indicator of tuning, as discussed in Section 7.

The second step is the computation of $\mu_h(e,e')$ from $\delta_h(e,e')$. The mapping must have negative derivative, but WRR's choice $\mu = 1/\delta$ is not the only possibility. Other possibilities are included in Table 6 (though the first is already in Table 5). Table 6 also shows the effect of slight changes to the definition of $H(d,d')$.

The practice of using the perturbed letter positions for measuring distances, introduced by WRR some time after the completion of the work reported in WRR94, has only a slight effect for both lists: $[\underline{0.8}, \underline{0.7}; 1.2, \underline{0.9}]$. Their other major change, replacing the definition of $\mathit{\Delta}(n,h)$ by one that is more geometrically correct, has a negligible effect.

The value $\sigma(e,e')$ is defined as a sum over h, but, as mentioned by WRR (1986), it could have been the maximum instead. That gives $[176, 6.3; 12.6, 3.9]$. If we are looking for the best term, we could also widen the search by including the values of h on each side of those in $H(d,d')$ $[280, 7.9; 26, 17]$, or two values on each side $[420, 11.2; 21, 15]$.

The definition of domain of minimality allows variation too. Instead of "smaller than d," we could use "smaller than or equal to d," or just take the whole text. Similarly, instead of using the size of the intersection to define the domain of simultaneous minimality, we could use the square of the intersection or other functions. Table 7 gives the scores.

Variation	Scores
Definition of $\mu_h(e,e')$*:*	
$1/\sqrt{\delta}$	[154, 120; 10.1, 99]
$1/\delta^2$	[560, 6.0; 26, 2.5]
$-\delta$	[5e8, 6100; 1e8, 7e5]
$-\delta^2$	[5e8, 2e4; 1e8, 7e5]
$-\ln\delta$	[6e8, 3000; 1e8, 8e5]
$\exp(-\delta)$	[3e6, 240; 250, 33]
Definition of $H(d,d')$*:*	
Round $\frac{1}{2}$ down	[1.1, 1.0; 1.4, 1.5]
Always round down	[$\underline{0.8}$, $\underline{0.8}$; 1.5, 1.6]
Always round up	[1.4, 1.0; $\underline{0.4}$, $\underline{0.6}$]
Remove duplicates	[$\underline{0.5}$, $\underline{0.7}$; 1.5, 1.7]
Use 1 value of i	[2e5, 340; 31, 21]
or 2	[2e4, 210; 3.4, 4.5]
or 5	[3.7, $\underline{0.6}$; $\underline{0.3}$, $\underline{0.2}$]
or 10 (WRR)	[**1**, **1**; **1**, **1**]
or 15	[3.6, 3.3; 1.4, 1.1]
or 20	[11.8, 5.9; 3.1, 3.8]
or 25	[66, 15.3; 4.8, 5.4]
or 50	[3600, 40; 93, 28]
Minimum row length 3	[$\underline{0.9}$, 1.0; 1.3, 1.2]
or 4	[$\underline{0.9}$, 1.0; 1.0, 1.1]
or 5	[$\underline{0.9}$, 1.0; 1.2, 1.3]
or 10	[1.1, $\underline{0.9}$; 5.4, 5.9]

Table 6. The effect of changing $\mu_h(e,e')$ or $H(d,d')$

Next consider the definition of the key function $\Omega(w,w')$. WRR defined it as a sum, but they could also have taken the best term $[4700, 13.6; 64, 1.8]$. If the best term is taken there, it makes sense to also take the best term in defining σ $[2e5, 12.5; 690, 10.2]$, perhaps with the search expanded to more h values, as described above: $[1e5, 23; 2200, 52]$ and $[9e4, 22; 2900, 100]$.

Another important part of the definition of $\Omega(w,w')$ is the definition of $E(w)$. WRR define it according to a skip limit with parameter 10 (an expected number of ELSs, as described before). The value 10 is not sacred; in fact, it is stated in WRR94 that a limit was only used to reduce the computational effort. However, as Table 8 shows, there is a clear optimum near 10 for both lists! (As an aside, we note that if we take WRR at their

Variation	Scores
Definition of T_e:	
Use $\leq$	[1.3, 1.1; 3.7, 2.7]
Whole text	[27, 850; 2.0, 407]
Definition of $L\omega(e, e')$:	
$\lvert T_e \cap T_{e'} \rvert^2$	[36, 1.5; 12.1, 1.1]
$\lvert T_e \cup T_{e'} \rvert$	[94, 580; $\underline{0.2}$, 29.1]
... but only if disjoint	[27, 52; $\underline{0.5}$, 19.0]
$\lvert T_e \rvert \, \lvert T_{e'} \rvert$	[4.6, 1.3; 2.2, $\underline{0.8}$]
$(\lvert T_e \rvert + \lvert T_{e'} \rvert)/2$	[4.8, 42; $\underline{0.5}$, 11.9]
$\sqrt{\lvert T_e \rvert \, \lvert T_{e'} \rvert}$	[2.7, 5.8; $\underline{0.8}$, 6.3]
$\min(\lvert T_e \rvert, \lvert T_{e'} \rvert)$	[1.1, 1.7; $\underline{0.9}$, 1.1]
$\max(\lvert T_e \rvert, \lvert T_{e'} \rvert)$	[109, 470; $\underline{0.4}$, 27]

TABLE 7. Various definitions of domains of minimality

word that the bound of 10 was only for computational efficiency, we must conclude that the "true" result of their experiment was one or two orders of magnitude weaker than claimed.)

Variation	Scores
Expected ELS count of 2	[7600, 7.0; $4e4$, 310]
or 5	[53, 1.6; 20, 19.5]
or 10 (WRR)	[**1**, **1**; **1**, **1**]
or 15	[1.2, 2.9; 5.9, 2.0]
or 20	[2.7, 8.3; 59, 7.1]
or 25	[$\underline{0.8}$, 4.0; 91, 15.2]
or 30	[6.8, 14.1; 144, 22]
or 50	[2.2, 4.1; 550, 79]
or 75	[3.7, 4.5; 590, 81]
or 100	[4.0, 4.7; 560, 62]
Exactly 10 ELSs	[23, 2.2; 630, 7.7]
Minimum skip of 1	[1.5, 2.1; $\underline{0.1}$, 5.0]
or 3	[$\underline{0.3}$, $\underline{0.7}$; 11.1, 5.9]
or 4	[1.2, 1.6; 16.3, 7.9]
or 5	[$\underline{0.5}$, $\underline{0.8}$; 16.7, 11.3]
or 10	[13.7, $\underline{0.6}$; 33, 35]

TABLE 8. The effect of changing $E(w)$

The sharp cut-off at parameter 10 allows us a simple experiment which to some extent is independent of the original experiment. We did the same computation restricted to those ELS pairs which lie within the cut-off at parameter 20 but not within the cut-off at parameter 10. Out of all eight statistics (P_{1-4} for each list), there is no value less than 0.418 and no permutation rank less than 0.342.

The use of the correct formula for defining $E(w)$ (see Section Appendix A), or whether the boundary is rounded up or down, have no effect (to the accuracy we are measuring it). However, some other variations do have an effect. Choosing the 10 ELSs with least skip, rather than all those within a boundary chosen to give 10 on average, affects the result a lot, as does using a lower bound other than 2 for the skip. These results, shown in Table 8, show that the result for the second list owes a lot to ELSs with very small skips, at which scales the strong nonrandomness of the text makes the method of perturbations nonsensical.

Next we consider the definition of the perturbations (x, y, z). Instead of applying them to the last three letters, we could follow the diagram given originally by WRR (1986) (but apparently not used in the calculations there) and apply them always to the third, fourth, and fifth letters, or we could apply them in pattern x, y, z instead of $x, x+y, x+y+z$. We could also try perturbing two letters instead of three, or perturbing them by larger amounts. Another variation in the use of perturbations, suggested by Witztum, is to only perturb the ELSs for the dates and use unperturbed ELSs for the appellations. We tried it the other way round as well. The scores for all these variations appear in Table 9. Note that using perturbation amounts as large as the skip is absurd, as different letters can be sought at the same position in the text. The two very small P_4 ratios (0.04 and 0.005) in the table are artifacts caused by that anomaly. Restricting the skip to be greater than the maximum perturbation increases them to 1.9 and 0.2, respectively.

Table 10 shows the effects of the lower bound 10 for the number of defined $\Omega^{(x,y,z)}(w, w')$ values, appearing in the definition of $c(w, w')$. The same table shows the effect of changing the cut-off 0.2 used to compute P_1 and P_3. Values greater than 0.2 have a dramatic effect on P_1, reducing it by a large factor (especially for the first list). However, the result of the permutation test on P_1 does not improve so much, and for the second list it is never better than that for P_4.

In applying the permutation test, there are a few more possible variations. Some rabbis have either no dates or no appellations in WRR's lists. In one case, they selected no appellations within their self-imposed length bounds of 5–8 letters. In other cases, they eliminated dates on the grounds that they were uncertain. Removing such rabbis has a minor effect, $[1.0, 1.2; 1.0, \underline{0.9}]$. In addition, some of the other rabbis produce no distances

Variation	Scores
Perturb as x, y, z	[$\underline{0.7}$, $\underline{0.1}$; $\underline{0.8}$, 2.1]
Perturb letters 3,4,5	[$\underline{0.4}$, 1.0; 1.3, 2.1]
Perturb up to 3 places	[$\underline{0.2}$, 2.4; $\underline{0.04}$, 1.1]
or 4 places	[$\underline{0.2}$, 4.2; $\underline{0.005}$, $\underline{0.6}$]
Perturb last 2 letters	[5e4, 4.5; 6700, 28]
up to 3 places	[118, 2.4; 340, 18.6]
or 4 places	[2.5, $\underline{0.6}$; 135, 48]
Perturb only appellations	[23, 7.5; 240, 34]
Perturb only dates	[15000, $\underline{0.3}$; 1350, 7.3]

TABLE 9. Different ways to do perturbations

Variation	Scores
Denominator bound:	
2	[2.9, 1.0; 1.0, 1.0]
3	[2.9, 1.2; 1.0, 1.0]
4	[1.8, 1.2; 1.0, 1.0]
5	[1.8, 1.2; 1.0, 1.0]
15	[1.0, 1.0; 1.0, 1.0]
20	[1.0, $\underline{0.9}$; 1.1, 1.1]
25	[1.0, 1.0; 1.1, 1.1]
Cut-off defining P_1*:*	
0.05	[**1**, 1.0; **1**, 1.0]
0.1	[**1**, 1.0; **1**, 1.0]
0.15	[**1**, 1.0; **1**, 1.0]
0.25	[**1**, $\underline{0.8}$; **1**, 1.0]
0.33	[**1**, 1.0; **1**, 1.0]
0.4	[**1**, 1.0; **1**, 1.0]
0.5	[**1**, $\underline{0.4}$; **1**, 1.0]

TABLE 10. Different denominator bounds or P_1 cut-offs

either (because of appellations or dates having no ELSs); removing all rabbis that produce no distances has the effect $[1.0, \underline{0.4}; 1.0, 7.8]$.

Acknowledgments

Many people contributed in substantial ways to this work. We particularly wish to thank David Assaf, Robert Aumann, Menachem Cohen, Persi Diaconis, Simcha Emanuel, Alec

Gindis, Michael Hasofer, Althea Katz, Mark Perakh, James Price, Barry Simon, Shlomo Sternberg, Jeffrey Tigay, Emanuel Tov, Ian Wanless, and the late Michael Weitzman. In addition, we owe a considerable debt to several people who donated large amounts of time and expertise but do not wish to be named.

We also thank WRR for providing us with some unpublished material.

Special thanks is due to the four referees, whose unusually thorough reports greatly improved this paper.

References

Many of the references are to documents published on the Internet. To simplify the citation, we will abbreviate two important sites as follows:

M = http://cs.anu.edu.au/~bdm/dilugim

W = http://www.torahcodes.co.il

All the references listed here as Internet documents can also be obtained from the authors.

Anonymous (1999). Equidistant letter sequences in Tolstoy's *"War and Peace"*: Witztum's "refutation" refuted. M/WNP/Rebuttal.pdf.

Bar-Hillel, M., Bar-Natan, D., and McKay, B. D. (1997). There are codes in *War and Peace too* (in Hebrew). *Galileo* **25** Nov-Dec 52–57.

Bar-Hillel, M., Bar-Natan, D. and McKay, B. D. (1998). Torah codes: puzzle and solution. *Chance* **11** 13–19.

Bar-Natan, D., Gindis, A. and McKay, B. D. (1999). Report on new ELS tests of Torah, revised report. M/report2.html.

Bar-Natan, D. and McKay, B. D. (1997). Equidistant letter sequences in Tolstoy's *"War and Peace"* (draft). M/WNP/draft.

Bar-Natan, D. and McKay, B. D. (1999). Equidistant letter sequences in Tolstoy's *"War and Peace"*. M/WNP.

Bar-Natan, D., McKay, B. D. and Sternberg, S. (1998). On the Witztum-Rips-Rosenberg sample of nations. M/Nations.

Bik, A. (1974). *Rabbi Yaacov Emdin* (in Hebrew). Mosad Harav Kook, Jerusalem.

Breuer, M. (1976). *The Aleppo Codex and the Accepted Text of the Bible*. Mosad Harav Kook, Jerusalem.

Cohen, M. (1997a). Two letters, M/cohen.html and M/WNP/CohenLetter2.html.

Cohen, M. (1997b). Personal communication.

Cohen, M. (1998). The idea of the sanctity of the Biblical text and the science of textual criticism. M/CohenArt. [Translated from the Hebrew original in *The Scriptures and Us*, Uriel Simon, ed. (1979). The Center for Judaism and Contemporary Thinking and Dvir, 42–69, Tel-Aviv.]

Cross, F. M. Jr. and Freedman, D. N. (1952). *Early Hebrew Orthography. A Study of the Epigraphic Evidence*. American Oriental Series **36**. American Oriental Society, New Haven.

Diaconis, P. (1986). Letter of December 30 to D. Kazhdan.

Diaconis, P. (1990). Letter of September 5 to R. Aumann. Reproduced with permission at M/witztum/PDtoRA.gif.

Dorfmann, D. D. (1978). The Cyril Burt question: new findings. *Science* **201** 1177–1186.

Drosnin, M. (1997). *The Bible Code*. Simon and Schuster, New York.

Em Piqchit (pseudonym) (1998). Chanukah candles in *War and Peace*. M/candles.

Encyclopedia Hebraica (1988). (in Hebrew) The Society for Publication of Encyclopedias, Jerusalem.

Encyclopedia Talmudica (1992). Entry "Iyyar". Talmudit Publishing, Jerusalem.

Fisher, R. A. (1965). *Experiments in Plant Hybridisation*. Oliver and Boyd, Edinburgh.

Freedman, D., Pisani, R. and Purves, R. (1978). *Statistics*. Norton, New York.

Gans, H. (ca. 1995). Coincidence of equidistant letter sequence pairs in the Book of Genesis. preprint.

Gans, H. (1998). Public statement of March 24. Available at M/Gans.html.

Gedaliah of Siemiatycze, (Rabbi) (1963). *Pray for the Peace of Jerusalem* 10–11 (in Hebrew). [This disagrees with *The Travels of Rabbi Moshe Yerushalmi*, Amsterdam 1769. Hebrew translation in A. Yaari (1946). *Travels in the Land of Israel*, 448, Tel-Aviv, Israel.]

Ha-Nagid, S. (1926). *Anthology of Poems of Shmuel Ha-Nagid* **1**, 1 (in Hebrew). Hebrew Union College Press, Cincinnati.

Hasofer, A. M. (1998). A statistical critique of the Witztum *et al.* paper. M/hasofer_s.pdf.

Havlin, S. Z. (1996) Statement of opinion. W/havlin.htm.

Kaczynski, T. (1995) Industrial society and its future (the Unabomber Manifesto). *Washington Post*, Sept. 19. Copy available at M/unabomber.txt.

Kahneman, D. and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology* **3** 430–454.

Kalai, G., McKay, B. D. and Bar-Hillel, M. (1998). The two famous rabbis experiments: how similar is too similar? Discussion paper 182, Center for Rationality and Interactive Decisions, The Hebrew University of Jerusalem.

Kass, R. (1994). Editor's remarks on WRR94. *Statist. Sci.* **9** 306.

Kass, R. (1998). Public statement at http://lib.stat.cmu.edu/~kass/biblecodes.

Khalifa, R. (1992). *Quran, The Final Testament*. Universal Unity, Fremont.

Mabit (16th century). *Responsa* I, 1G 103 (in Hebrew).

Margaliot, M. (ed.) (1962). *Encyclopedia of Great Men of Israel* (in Hebrew). Joshua Chachik, Tel-Aviv.

McCormack, R. (1923). *The Heptadic Structure of Scripture*. Marshall Brothers, London.

McKay, B. D. (1999a). In search of mathematical miracles. M/index.html.

McKay, B. D. (1999b). Torah codes. M/torah.html.

McKay, B. D. *et al.* (1998). Jesus as the Son of Man. M/Jesus.

McKay, B. D. (1998). An objective experiment of Doron Witztum. M/witztum/camps.html.

Naveh, J. (1987). *Early History of the Alphabet: An Introduction to West Semitic Epigraphy and Palaeography*. Magnes Press, Jerusalem.

Panin, I. (ca. 1908). *Verbal Inspiration of the Bible Scientifically Demonstrated*. Privately published.

Perakh, M. (1998). Various articles on Bible codes. http://www.bigfoot.com/~perakh/fcodes.

Reiner, R. (1997). *Rabbenu Tam* (in Hebrew). Masters Thesis, The Hebrew University of Jerusalem.

Rips, E. (ca. 1985). Transcript of lecture. M/ripslect.

Rosenberg, Y. (undated). Two programs, `els1.c` and `els2.c`, implementing versions of the function $c(w, w')$.

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research* (enlarged ed.). Irvington, New York.

Satinover, J. (1997). *Cracking the Bible Code*. Morrow, New York.

Schacter, J. J. (1988). *Rabbi Jacob Emden: Life and Major Works*. 21. PhD Thesis, Harvard Univ.

Simon, B. (1998). Various articles on Bible codes. http://wopr.com/biblecodes.

Teomim, Y. (1993). *New Interpretations by the Author of "Peri Megadim"* (in Hebrew). 315–316. Machon Sha'ar HaMishpat, Jerusalem. [Also see *Anthology on Mosaic Law* (1988). Kollel Torat Moshe al-shem HaChatam Sofer, 64 (in Hebrew).]

Thomas, D. E. (1997). Hidden messages and the Bible code. *The Skeptical Enquirer* **21** Nov-Dec 30–36.

Tigay, J. (1998). The Bible "codes": A textual perspective. http://www.sas.upenn.edu/~jtigay/codetext.html.

Tov, E. (1992). *Textual Criticism of the Hebrew Bible*. Fortress Press, Minneapolis.

Tov, E. (1998). Personal communication. (Tov is Magnes Professor of Bible at The Hebrew University of Jerusalem, and the co-editor-in-chief of the official Dead Sea Scrolls project.)

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bull.* **2** 105–110.

Ulrich, E., Cross, F. M., Davila, J. R., Jastram, N., Sanderson, J. E., Tov, E. and Strugnell, J. (1994). Qumran Cave 4, VII, Genesis to Numbers. *Discoveries in the Judaean Desert* **12**. Clarendon Press, Oxford.

Vilenski, E. (1949). Biography of Rabbi Immanuel Chai Ricchi (in Hebrew). *Kiriat-Sefer* **25** 311.

Williamson Hoke, K. (1988). Completely unimodal numberings of a simple polytope. *Discrete Appl. Math.* **20** 69–81.

Witztum, D. (1989). *The Added Dimension* (in Hebrew). Privately published.

Witztum, D. (1998a). A refutation refuted. W/ref1.htm, W/ref2.htm.

Witztum. D. (1998b). The seal of G-d is truth. W/emet_hb.htm.

Witztum, D., Rips, E. and Rosenberg, Y. (1986). Equidistant letter sequences in the Book of Genesis. Preprint.

Witztum, D., Rips, E. and Rosenberg, Y. (1987). Equidistant letter sequences in the Book of Genesis. Preprint.

Witztum, D., Rips, E. and Rosenberg, Y. (1994). Equidistant letter sequences in the Book of Genesis. *Statist. Sci.* **9** 429–438.

Witztum, D., Rips, E. and Rosenberg, Y. (ca. 1995). Equidistant letter sequences in the Book of Genesis, II. The relation to the text. Preprint. Copy at M/Nations/WRR2.

Witztum, D., Rips, E. and Rosenberg, Y. (1996). Hidden codes in equidistant letter sequences in the Book of Genesis, The statistical significance of the phenomenon (in Hebrew). Preprint.

Young, J. (1997). *Behold Yeshua!* For His Glory, Richardson.

Zevit, Z. (1980). *Matres Lectionis in Ancient Hebrew Epigraphs.* American Schools of Oriental Research Monograph, Ser. 2, American Schools of Oriental Research, Cambridge.

Ziegler, G. M. (1995). *Lectures on Polytopes.* Springer, New York.

Department of Computer Science, The Australian National University, Canberra, ACT 0200, Australia

E-mail address: bdm@cs.anu.edu.au

Department of Mathematics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

E-mail address: drorbn@math.huji.ac.il

Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem 91905, Israel

E-mail address: msmaya@math.huji.ac.il

Department of Mathematics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

E-mail address: kalai@math.huji.ac.il