THE TWO FAMOUS RABBIS EXPERIMENTS: HOW SIMILAR IS TOO SIMILAR?

GIL KALAI, BRENDAN MCKAY, AND MAYA BAR-HILLEL

Abstract. Witztum, Rips and Rosenberg [9] describe the outcomes of two experiments which purport to statistically prove the existence of a hidden code in the Book of Genesis. We show that these two experiments, viewed as two random samples from the same population, yielded numerical outcomes which are more similar to each other than expected. We also show that the distributions obtained in some control experiments performed by Witztum et al. are flatter than expected. Our hypothesis is that Witztum et al. tailored their experimental procedures to meet naive expectations regarding how outcomes of experimental replication and experimental controls should look. We give some statistical and empirical evidence supporting this hypothesis.

Contents

1. Introduction 2
2. The present hypothesis 3
3. Similarity of the outcomes of the two experiments 3
3.1. The p-values 3
3.2. The distance distributions 4
4. How did it happen? 4
4.1. The p-values: a process model 4
4.2. The distance distributions: a motive 5
5. Changes in the pairwise distances 7
6. Intuitive judgement 7
6.1. The p-value 7
6.2. The distance distributions 8
7. Uniformity in control experiments 8
8. Gans' experiment 9
9. Conclusion 10
Acknowledgments 10
References 10

1. Introduction

Witztum, Rips and Rosenberg (WRR, for short) describe two experiments they carried out on the Book of Genesis, which they claim to have statistically proven the existence of a hidden code in that book [9]. This code consists of words appearing in the form of equidistant letter sequences (ELSs, for short). An ELS is a sequence of letters chosen from the text at equal spacing. (Spaces between words, and punctuation, are ignored.) WRR defined a *pairwise distance* function $c(a, b)$ which, in some sense, measures the proximity of the ELSs of word a to the ELSs of word b. The value of $c(a, b)$ may be undefined, or may be a fraction between 1/125 and 1. Smaller values of c(a,b) are taken as meaning that the words a and b are "closer".

WRR's first experiment (1986) consisted of pairing 34 famous Jewish rabbis with their known dates of death or birth. These rabbis lived between the 7th and the 18th century (millennia after Genesis was written), and were known by many different names and appellations (much as Lady Diana was known as Diana Spencer, Princess of Wales, Lady Di, Diana, etc.) Their dates were written in three different forms (roughly corresponding to "July 4", "4th of July", and "on July 4", but bearing in mind that Hebrew dates can be written using only letters and no numerals). Pairing all the different names with all the dates for each rabbi yielded a total of 152 word pairs for which the pairwise distances were defined. WRR's "hidden code" hypothesis suggested that these distances should be inordinately small, rather than uniformly distributed as they imagined would happen by chance. Their calculations showed that the pairwise distances were indeed smaller than what they expected by chance, to a highly significant level.

Sometime in 1986, Harvard's Persi Diaconis was asked to evaluate this work. He suggested that WRR run a second test on a fresh sample, and that they compare its performance with a control test based on a random cyclic pairing of the rabbis with dates of other rabbis. In the latter case, of course, no effect is expected. WRR's second experiment consisted of 32 new rabbis, who yielded a total of 163 pairwise distances between these rabbis' various names and appellations and their respective date forms. The results of the new sample were strikingly similar to those of the first. The result of the control test showed no effect. These two experiments are sometimes referred to as "the famous rabbis experiments". Details of the experiments, as well as the actual lists, can be found in [9].

Even more details can be found in a 1987 preprint [8], which tabulates all 315 $c(a, b)$ values. By contrast, [9] gives the pairwise distances only in the form of two histograms, corresponding to the two lists. Both histograms show a strong skew towards small $c(a, b)$ values (see Figure 1).

`---Figure 1 about here---`

Skeptics have always supposed that WRR's amazing results are the outcome of an intentional or unintentional optimization process in the selection of the measurement tools or in the selection of the data themselves. A major turning point came when a search of the rabbinical literature, and of Bar-Ilan University's "Responsa" database, uncovered the fact that the 66 rabbis in WRR's lists were known by a total of at least twice as many names and appellations as were actually included in the lists. McKay, Bar-Natan and Bar-Hillel (MBB, for short) [6] showed that WRR's guidelines governing the first list were lax enough to enable the selection of a second list of appellations to succeed spectacularly on "War and Peace", a text on which no effect was expected. MBB also showed that WRR's lists exhibited many traces of bias in their data selection.

The present analysis complements that of MBB from the same skeptical point of view. Unlike that paper, however, it does not deal with the mathematical, grammatical, or historical choices made by WRR, and for the most part requires no computations on the Book of Genesis or on any other text. It uses mainly the 152+163 pairwise distances given in [8] (and available for downloading from [12]). These numbers are in agreement with the histograms in [9] except for a few minor changes.

2. The present hypothesis

Long before the Statistical Science paper was published, the results of the famous rabbis experiments were being touted, primarily in public lectures, but also in print (see, e.g., [10]). In these forums, the results were typically summarized two ways: by the alleged statistical significance values of the two lists and by the histograms. The prominence of these measures thus makes them salient ones.

Witztum and Rips themselves pointed out that the outcomes of the two experiments were very similar. Indeed, the two p-values for the statistical significance of the two experiments which appear in [8] are remarkably close and eyeballing the two histograms for the Genesis results (see Figure 1) also shows them to be quite similar.

Having similar outcomes for two experiments purporting to measure the same phenomenon is welcome. But when two measurements are much closer than the size of the measurement error allows us to expect, it might be suspiciously good—too good to be true, if you will. Our suspicion that the two p-values were closer than normatively expected arose after WRR on several occasions made minor corrections in their lists, which resulted in dramatic changes in the p-value. This suggested that it is a volatile measure.

We decided to subject our suspicion to a quantitative test. We tested the null hypothesis that the two p-values were as close as they were due to mere coincidence. The alternative hypothesis was that the closeness was the intended consequence of a flawed experimental procedure. Later we will put forth a concrete model for such a procedure, as well as possible motivation for using it. We will show that our model is compatible with what has been previously conjectured regarding WRR's data-selection process [6]. In addition, we will show excessive similarity to a uniform distribution in the outcome of the control test suggested by Diaconis. Finally, we will show that another famous rabbis experiment, reported by Gans [4], also came out "too good to be true".

3. Similarity of the outcomes of the two experiments

3.1. **The p-values.** A principal measure of statistical significance used by WRR in [8, 10] is their P_2 score: the probability that the product of uniformly distributed independent random variables is smaller than the product of the numbers observed in the experiment. After Diaconis pointed out that the assumption of uniformity and independence does not hold, WRR adopted a different p-value [9] (albeit one still based on P_2). For our purpose, the P_2 value remains pertinent, since any bias in the data selection occurred at a time when it served as the principal measure of significance.

The p-values corresponding to the P_2 scores of the two experiments given in [8] are 1.29×10^{-9} and 1.15×10^{-9}, respectively. The ratio between them is 1.12. We will presently see that in view of the instability of the P_2-statistic, such a small ratio is quite surprising.

Since we don't have a sampling distribution for this ratio, we assessed the probability of finding a ratio this close to 1 by the following procedure. Suppose we lump all 66 rabbis

from the two lists together, and then repartition them randomly into 34 and 32 rabbis. For each such partition, we can calculate the two P_2 scores and the ratio of the larger P_2 score to the smaller P_2 score. In a Monte Carlo simulation of the sampling distribution of the P_2 ratio, we found that less than one ratio in a hundred was 1.12 or smaller.

Is it legitimate to compare the original partition of 34 and 32 rabbis to random partitions? We claim it is, inasmuch as this comparison is biased, if at all, against us. The first list of rabbis was drawn from rabbis who have upwards of 3 columns of text in the Encyclopedia of Great Men of Israel [5], and the second list was drawn from rabbis who have between 1.5 and 3 columns of text. These two sets of rabbis differ in terms of their fame and greatness, which was the reason for looking at rabbis in the first place. A random partition of the 66 rabbis creates two sets with more variance within, but less variance between, than the original partition. Hence the P_2 ratio should, if anything, be *larger* for the original partition than for random ones.

To get a feeling for the variance of the distribution of the P_2-ratio for random partitions, note that the median value of this ratio is around 700. The distribution of the logarithm of the ratio is close to uniform below the median. This allows another, though cruder, estimate for the probability of obtaining such a small ratio: $2\log 700/\log 1.12 = 0.0088$ much like the Monte Carlo estimate (0.0092).

3.2. **The distance distributions.** At the heart of each of WRR's two experiments were the pairwise distances for all the pairs of appellations versus dates. The histograms, of course, are derived from these distances. So, we examined the similarity of the distributions of the pairwise distances between the two experiments. A common measure of distance between two distributions is the supremum norm of their difference (Kolmogorov-Smirnov distance) [11, Ch. 14]. The sup-norm distance between the two distributions of pairwise distances in the original partition is 0.05489. Figure 2 shows the cumulative pairwise distance distributions of the two experiments.

`--Figure 2 about here----`

In a Monte Carlo simulation, we compared the proximity between the distributions of the pairwise distances for the original partition of 66 rabbis with those for random partitions into 34 and 32. The Monte Carlo derived probability that the sup-norm distance between the two distributions of pairwise distances will be smaller than or equal to 0.05489 is 0.035.

4. How did it happen?

Having shown that the similarity of the two P_2 scores, and of the two pairwise distances distributions, are extraordinary (in the sense that rarely are such similarities encountered by chance), we are now prepared to suggest how this similarity came about. Since the definitions of the P_2 score and $c(a, b)$ function were already fixed by the first experiment, we need to account for the similarity by a process involving only the second experiment. Our explanations will attribute to WRR practices from which these similarities follow either as byproducts, or as a consequence of direct intention.

4.1. **The p-values: a process model.** Recall that MBB have shown that the second list was not really drawn up in quite as rigorous and objective a manner as WRR would have us believe. Rather, some manner of judgement or discretion was exercised when deciding which names and appellations to include in the second list, and which to reject ([6] Section 5.4). Of course, there wasn't complete freedom of choice with regard to the selection of names and appellations. Some names are "compulsory": they are so closely identified with

a rabbi that their absence would be noticed at once. The flexibility therefore applies only to the subset of "elective" names and appellations[1].

Our model supposes that the elective names and appellations are considered one by one, with those contributing to a small P_2 value added, and those increasing it deleted, until the P_2 level of the first experiment is first surpassed. We will now show that the observed P_2-ratio is compatible with several possible variations of this process. By "compatible" we mean that there is a reasonably large probability of obtaining the results actually observed. We analyzed two such variations, and simulated some others.

1. Consider a process whereby one starts with the minimal set of compulsory appellations, and from among the elective appellations one only adds favorable appellations, in a *random* order. For every appellation i in the second list, let x_i be its *effect* on the P_2 score. By the effect of an appellation a we mean the P_2 score of all the pairwise distances (163 in our case) divided by the P_2 score of all the pairwise distances except those yielded by a. For convenience we will discuss the log effects. It is reasonable to assume that the probability that the critical appellation was i is proportional to $\log x_i$ and is 0 when $\log x_i < 0$. To see this, reverse the order so that first the process is carried out for *all* the elective appellations, and only later the critical value is determined (at random), see Figure 3.

`---Figure 3 about here---`

The probability of a log P_2-ratio which is as small as $\log 1.12$ is 1 if $\log x_c \leq \log 1.12$ and is $\log 1.12/\log x_c$ if $\log x_c > \log 1.12$, where c is the critical appellation. Based on these assumptions and a computation of the effect of all individual appellations, we estimated the percentile of 1.12 in the distribution of P_2-ratios yielded by such a process to be 19.

2. Consider a process whereby one starts with all known appellations and successively deletes elective unfavorable appellations, in random order. Using the same computation, and basing it on the appellations with the negative effect, we estimate the percentile of the observed P_2 ratio, 1.12, to be 30.

3. In the context of WRR's experiment it is reasonable to suppose that appellations with larger impact were treated earlier, rather than in random order. We considered especially a model where unfavorable elective appellations were deleted in the order of their P_2-values. For simulations based on this model, the percentile of a P_2-ratio of 1.12 ranged between 40 and 70, depending on various parameters of the process. It is interesting to note that the limit distributions of pairwise distances we obtained from these simulations were rather close (in the sup-norm sense) to those of WRR's experiments.

4.2. **The distance distributions: a motive.** The fact that the two lists of pairwise distances from the two experiments are as similar as they are was no doubt welcomed by WRR, but risked arousing some unease if we take into account the major volatility and non-robustness of WRR's phenomenon (see [6]). We could think of no simple process model, of the kind just presented, that could give rise to such a similarity. If WRR used similar optimization processes in selecting the data for the two experiments, this can explain why the two lists of pairwise distances look like two samples from the same distribution, but not why they are even more similar than most pairs of samples from the same distribution.

Could there have been, in addition to biased selection, also some intervention in the $c(a, b)$ values, to foster the similarity of list 2 to list 1?

[1]We neither want to, nor can, clearly separate the set of available names and appellations into "compulsory" and "elective".

Since the pairwise distances were typically given to audiences (including the later audience of Statistical Science readers) by the histograms summarizing them, we tested the possibility of intervention indirectly by checking for signs of intervention towards similar histograms. Our measure of proximity between two distributions, namely their sup-norm distance, is independent of this particular way of presentation. Suppose the similarity between the two $c(a,b)$ lists, however striking it may or may not be, was arrived at by mere chance. Then the similarity between various histograms representing these two lists, respectively, should not depend on particulars of the histogram display. In particular, the actual histograms displayed in [9] should not enjoy any systematic advantage over other similar histograms. We tested whether this was so.

Recall that WRR chose to display the pairwise distances as 25-bin histograms, where the i-th bin corresponds to pairwise distances in the interval $(0.04(i-1), 0.04i]$. (Thus the second bin, for example, corresponds to all pairwise distances which are bigger than 0.04 and smaller than or equal to 0.08.)

Let $H_a = (a_1, a_2, \ldots, a_{25})$ and $H_b = (b_1, b_2, \ldots, b_{25})$ be the vectors represented by WRR's two histograms. Thus, a_i, b_i $1 \le i \le 25$ are the number of pairwise distances in the i-th bin for the first and second experiments, respectively. Let $c_i = a_i + b_i$, let $r_a = \sum a_i / (\sum c_i)$ $(= 153/315)$, and let $r_b = 1 - r_a$. A standard measure of the proximity between the two pairwise distance distributions in terms of their histograms is the "binomial homogeneity test" [3], which asymptotically behaves (in the case of independent samples) like χ^2 with 24 degrees of freedom, and is given by:

$$D(H_a, H_b) = \sum_{i=1}^{25} ((a_i - r_a c_i)^2 / r_a c_i \ + \ (b_i - r_b c_i)^2 / r_b c_i).$$

The value of $D(H_a, H_b)$ for histograms based on the pairwise distances tables of WRR's two experiments is 16.15[2]. In order to study the dependence of this similarity measure on the particulars of the histograms, holding the $c(a,b)$ values constant, we introduced a small shift r, $-0.02 \le r \le +0.02$, and put the pairwise distance data into the 25 bins of the histograms, but based on the artificial "intervals" $(r + 0.04(i-1), r + 0.04i]$ (modulo 1). (For $r \neq 0$, one of these "intervals" is actually a union of the two endpoint intervals, one containing 0 and the other containing 1.) Of course, the histograms WRR chose are the only natural ones with 25 bins. Our shift is just a mathematical device to study the proximity between the histograms.

Let $H_a[r] = (a_1[r], a_2[r], \ldots, a_{25}[r])$ and $H_b = (b_1[r], b_2[r], \ldots, b_{25}[r])$, where $a_i[r]$, $1 \le i \le 25$, is the number of pairwise distances of the first experiment in the interval $(r + 0.04(i-1), r + 0.04i]$ (modulo 1) and $b_i[r]$ is the number of pairwise distances of the second experiment in that interval. We calculated the values of $D(H_a[r], H_b[r])$ for all histogram shifts r between -0.02 and $+0.02$. We were interested in checking for which histogram shifts r, $D(H_a[r], H_b[r])$ is smaller than or equal to to 16.15. It turned out to be true for all of them, since the minimum of the function $D(H_a[r], H_b[r])$ is attained for $r = 0$, which is the original presentation. Moreover, the values of r for which the minimum is attained consist of the very small subinterval $[0, 0.00034]$ of $[-0.02, +0.02]$. See Figure 4.

`---Figure 4 about here---`

Thus, we see that the natural histograms exploit the similarity between the distributions to the maximum possible. No other 25-bin histograms presenting the same two sets of

[2]Recall that there were minor differences between the tables in [8], the histograms in [8] and the histograms from [9]. Our analysis requires the full tables.

numbers are quite as similar. We are not saying that WRR chose their histograms from among the ones we just studied in order to maximize their similarity. Clearly theirs was the only natural choice. Rather, we are saying that the striking advantage of their histograms over shifted histograms suggests intervention motivated towards increasing these histograms' visual similarity.

Note that by itself the similarity of the two histograms in terms of $D(H_a, H_b)$, 16.15, is not very impressive. Two histograms H'_a and H'_b obtained by randomly partitioning the 66 rabbis into a set of 34 and a set of 32 and presenting the pairwise distances of the two parts will satisfy $D(H'_a, H'_b) \leq 16.15$ with probability 0.07.

5. Changes in the pairwise distances

WRR provide their own computer program `els1.c` for computing pairwise distances $c(a, b)$. To our surprise, the values differed in many instances from the values listed in [8]. Consequently it gave different histograms as well. It turned out that, while the numerical analysis in [9] was based on `els1.c`, the histograms (for Genesis), although presented in the 1994 paper, were actually based on defunct programs which WRR claim no longer exist.

`---Figure 5 about here-----`

Comparing Figure 1 to Figure 5, we observe that the two versions differ for 28 distances in the first experiment and 42 distances in the second. The excessive similarity between the two distributions disappears for the outcomes of `els1.c`. The sup-norm distance between the two distributions and the distance between the histograms are both quite close to the median values for random partitions of the set of 66 rabbis.

In all cases, we have used the distances in [8] for those experiments in this paper which only needed those distances, and `els1.c` for experiments needing other distances.

6. Intuitive judgement

What could WRR's motivation have been to achieve a replication so similar to the original experiment as to arouse suspicion? Possibly, WRR did not realize that their optimizations were leaving telltale signs, and thought that the results they showed for the second experiment were just what one should have expected, rather than better than expected. Such a view is congruent with findings about peoples' intuitive judgment under uncertainty in the psychological literature (see, e.g., [1])

6.1. **The p-value.**

Tversky and Kahneman studied people as intuitive statisticians. One of their studies concerned scientists' conceptions of replication [7]. They showed that if an experimental sample is characterized only in terms of the statistical significance of its results, and then a replication is planned, people have inflated intuitive expectations of achieving the same significance in the replication. In one of their questions they presented a scenario where, because of a smaller sample size, the replication had only 50% probability of yielding a significant result, yet the median respondent thought the probability was about 85%. In another scenario people regarded a smaller significance value in a second experiment as a failure to replicate the first one, although this smaller significance was not surprising even given the effect in light of the sampling error in that scenario. The respondents to these questions were statistically savvy mathematical psychologists.

When WRR aimed to replicate the statistical significance of their first list, they may have been unaware that the resulting 12% gap is unreasonably small. Indeed, realizing that this gap is too small requires considerable analysis.

6.2. **The distance distributions.** Tversky and Kahneman [2] also showed that people often expect even small samples to resemble their parent population (hence each other), more than they typically do. We decided to check peoples' perceptions in the present case directly.

Forty four persons (35 undergraduate students and 9 of their teachers in the Department of Computer Science at the Australian National University) were given a sheet of paper on which they saw two pairs of histograms: those from Figure 1 and Figure 5. They read the following question:

"There is a large barrel filled with millions of colored balls, of 25 different colors. Take out about 150 balls at random and draw a histogram of the number of balls of each color. Then take out a similar number of balls at random again and draw the histogram for them. Now you have two histograms. Below are two pairs of histograms. Which pair best matches your intuition of what your two histograms might look like?"

The published histograms were preferred to those generated by `els1.c`, by about 70% of the respondents. The students and teachers answered similarly. When we ran the same experiment on 53 students at Yale University, about 50% of the respondents preferred each one of the two pairs of histograms. These experiments tend to confirm that the published defunct histograms in [9] indeed meet peoples' expectations but, are inconclusive as to whether they are closer to peoples' expectations than the `els1.c` histograms.

7. Uniformity in control experiments

Recall that Diaconis had asked WRR to test a random cyclic pairing of rabbis with dates of different rabbis. The results of this control test were not presented in [9], but they were in [8]. WRR chose a particular cyclic shift: pairing rabbi i to the dates of rabbi $i+1$ (modulo 32)[3]. We refer to this control experiment as A. WRR based their early statistical analyses on the assumption that without the effect of the hidden code, the pairwise distances are uniformly distributed. Therefore they expected this control experiment to yield a uniform pairwise-distance distribution. The 25-bin histogram for the pairwise distances of A (Figure 6(a)) is indeed quite flat.

To see what distributions cyclic date shifts really produce, we considered the combined distribution for all cyclic shifts, namely all the pairwise distances of rabbi i versus dates of rabbi j for $j \neq i$. Figure 7 shows an unexpected skew towards small pairwise distances.

`---Figures 6 and 7 about here----`

What can explain this skew? The $c(a,b)$ function contains elements that depend not only on how close a is to b, but also on qualities of a and b in themselves. Therefore, if a fixed word a generates a small $c(a,b)$ with *some* word b it will tend to generate smaller $c(a,b)$ with *any* word b. WRR refer to such words as *charismatic*. The skeptics' hypothesis of biased appellation-selection suggests that WRR selection process favored charismatic appellations. Given this observation, it is not surprising that Figure 7 is not flat; rather it is surprising that WRR's histogram for A *does* look flat.

We used the χ^2 statistic of the histogram (relative to the uniform distribution) as a measure of flatness. Histogram A was flatter than the histogram for any other cyclic shift. To estimate the probability for a histogram as flat as or flatter than A using a Monte-Carlo simulation, we sampled a large number of (general) permutations without

[3]This choice involved a technical difficulty, since in a few cases two different rabbis were both called "Rabbi so-and-so". Rather than avoid the few cyclic shifts where this problem arises, WRR solved it by omitting *all* appellations of the form "Rabbi X".

fixed points. The probability for a random permutation without fixed points to generate a sample with a χ^2 value equal to or smaller than A's is 0.003[4].

The preprint [8] contains histograms for three additional control tests (denoted by S, M and R) in which the first list was run on three different texts (see Figure 6(b,c,d)). The histograms for S and M are surprisingly flat: the probabilities for histograms of the uniform distribution being as flat or flatter according to χ^2 are 0.004 and 0.015, respectively.[5] In the case of M, it turned out that the histogram in [8] is flatter (by about 5 times on the same scale) than the data in [8] actually gives. In the case of S, the Samaritan version of Genesis, our requests to WRR for the text were not successful, and the closest we could obtain (another edition) does not give a flat histogram at all.

To show that WRR's overly flat histograms are not intuitively perceived as overly flat, the same respondents mentioned earlier were also given eight histograms, the four histograms for the control tests from [8] and four random histograms of the same size. They read the following question:

"Now suppose the number of balls of each color in the barrel is exactly the same. Take out 100-150 balls at random and draw a histogram of the number of balls of each color. Below are eight histograms. Mark which four best match your intuition of what your histogram might look like?"

About 70% of the respondents included A in their choice set, and a similar percentage included the histograms for S and M. R was chosen by only about 25%. Altogether A, M, S and R were chosen 104 times as compared to 62 times for the other histograms. This time, the Yale students answered similarly.

8. Gans' experiment

More recently, another famous rabbis experiment was conducted [4]. WRR's 66 rabbis were paired with their places of birth and/or death rather than with their dates. A very low p-value was reported by Gans, and is commonly cited as evidence for the integrity of WRR's list of appellations. Skeptics will note, however, that Gans' experiment also suffered from too much room for data selection. Gans, who does not speak Hebrew, received the various names and spellings for the places from a colleague of Witztum. He developed his own measure of proximity, but it is strongly correlated with WRR's function $c(a, b)$. His overall p-value was based on a permutation test using his new measure.

This gave us an opportunity to check whether this new data also showed evidence of having been selected to generate overly similar experimental results for the two lists.

The p-values for the two lists that Gans analyzed were less than 2% apart, even smaller than the gap for WRR's p-values. We thank Dror Bar-Natan for computing these p-values and for doing the Monte Carlo simulation which provided an estimate of the probability that a random partition of the 66 rabbis into sets of 34 and 32 would yield two p-values as close as that or closer. This probability was around 1/500. Bar-Natan's p-values differ from Gans', because his analysis was based on the assumption that insofar as data selection occurred (and MBB gave evidence that it did), it was geared towards WRR's proximity measure and not Gans'. He used the permutation rank of P_2 rather than the value of P_2 for the same reason: it was the measure of success employed at the time by WRR.

[4]If we follow WRR's decision to exclude the "Rabbi X" appellations the probability for such a flat histogram is 0.02.

[5]Although the observations here are not independent, tests we did with a large number of artificial texts failed to detect any general trend towards flatness.

9. Conclusion

WRR's fantastic claims raise the question whether the outcomes they describe express their own wishes rather than any real phenomenon. This paper claims that WRR's results stretch credibility, even without challenging the validity of their hidden code hypothesis. Our analysis of the results of their replication and control experiments show them to express naive expectations rather than statistical reality.

Note that our case is presented as a whole, and does not rest on any particular claim. Indeed, some of the phenomena we observed may well have been the result of chance or of some indirect mechanism we have not identified. However, the combined weight of the evidence appears to us to be considerable.

Acknowledgments. It is a pleasure to acknowledge the great help by Danny Braniss, Leo Novik and Michael Ukon in programming. We would like to thank also Dror Bar-Natan, Yosi Rinnot, Boris Tsilerson and Ilya Rips for helpful discussions. Finally, we would like to mention the early work of Nahman Givoli who pioneered the critical study of ELS claims.

References

[1] D. Kahneman, P. Slovic and A. Tversky (eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, 1982.
[2] D. Kahneman and A. Tversky, Subjective probability: A judgment of representativeness, Cognitive Psychology, 3(1972) 430-454.
[3] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, 3rd Edition, 1973, Vol. 2, page 598.
[4] H. Gans, Coincidence of Equidistant Letter Sequence Pairs in the Book of Genesis, preprint, ca. 1995.
[5] M. Margalioth (ed.), Encyclopedia of Great Men of Israel (Hebrew), Joshua Chachik, 1962.
[6] B. McKay, D. Bar-Natan, and M. Bar-Hillel, The Bible Code: The genesis of equidistant letter sequences, submitted.
[7] A. Tversky and D. Kahneman, Belief in the law of small numbers, Psychological Bull. 2 (1971), 105-110.
[8] D. Witztum, E. Rips and Y. Rosenberg, Equidistant letter sequences in the Book of Genesis, preprint, 1987.
[9] D. Witztum, E. Rips and Y. Rosenberg, Equidistant letter sequences in the Book of Genesis, Statistical Science, 9 (1994), 429-438.
[10] D. Witztum, The Additional Dimension, Jerusalem, 1992.
[11] S. Wilks, *Mathematical Statistics* Ch. 14, Wiley, New-York, 1962.
[12] http://sunset.huji.ac.il/ kalai/bc

Institute of Mathematics, Hebrew University, Jerusalem 91904 Israel.
E-mail address: kalai@math.huji.ac.il

Department of Computer Science, The Australian National University
E-mail address: bdm@cs.anu.edu.au

Center for the Study of Rationality, The Hebrew University of Jerusalem
E-mail address: msmaya@olive.mscc.huji.ac.il