A Primer on the Torah Codes Controversy for Laymen

Harold J. Gans

CONTENTS

Harold Gans was a senior Cryptologic Mathematician with the National Security Agency, US Department of Defense, for 28 years until his retirement in 1996.

Foreword

By Professor Robert Haralick
Boeing Clairmont Egtvedt Professor of Electrical Engineering
University of Washington

One of the most important elements in participating in a dialog on the Torah Code Controversy is clarity. We have to know what the Torah Code hypothesis is and what it is not. And we have to be able to separate any assumptions arising from our personal religious beliefs from the logic of the statistical debate. This is because the debate is not a religious debate. The Torah Code hypothesis is a hypothesis in the statistical domain about the Torah text that we have today. Its language is the language of probability and statistics.

The debate is over the different experiments that have been performed and whether or not the experimental results that have been observed can be explained in "natural terms" or not. One side of the debate holds that the experiments are proper experiments and there is no natural explanation for the observed results. And hence the Torah Codes are real. The other side of the debate holds that there is indeed a natural explanation: simply that the experiments were not proper a priori experiments and, therefore, the Torah Codes are not real.

The politics of the debate has religious and secular elements opposed. Underlying the discussions are often strong emotions and this makes it more difficult to be logical and to be clear. And confusing the matter even more is the fact that the debate takes place in an environment in which there is a tabloid-like layer that is certainly orthogonal to the rigors of the statistical debate. This tabloid-like layer, due to popular books, from religious (Jewish and non-Jewish) and non religious people provide a variety of Torah Code arrays showing remarkable closeness between ELSs of historically or logically related key words. However, all these kinds of examples must be regarded as either meaningless or anecdotal, because they were not generated in accordance with a proper a priori experimental protocol in which first the related key words are specified and second an experiment is run which determines a probability that more ELSs of the related keywords are "closer together" than expected by chance, whatever chance might mean.

Further complicating the issue is that in this situation "chance" does not have a unique meaning. To estimate a probability requires that a control population be specified. The control population can be specified as a set of randomized texts or a set of randomized pairing of key words, where randomized itself can have a variety of different possible meanings. The important point about the control population is that it is designed not to have very many significant events. Change the population and we change the probability.

Then there is the issue of the way in which the observed experimental data are combined or analyzed to generate the test statistic which is compared from experimental trial to trial. Finally, there is the issue of the definition of ELS closeness or compactness.

There are a variety of different measures that can be used, each yielding different experimental results.

By random sampling the events from the control population and observing how many of the sampled events from the control population have a smaller valued test statistic than that observed with the Torah text and the correct pairing of related key words, we can estimate a probability of observing as significant or more event in the control population. This is the p-value associated with the experiment. When the p-value is sufficiently small, we declare that the observed results are not due to chance.

If the Torah Codes do not exist, neither the definition of control population, nor the definition of compactness measure, or the difference in test statistic will make any difference. However, if the Torah Codes are real, then each of the choices for control population, test statistic, and compactness measure can make substantial differences in the effectiveness of the technique in detecting the existence of Torah Codes. This is because for each experimental combination there will be a false alarm probability and a misdetect probability. The false alarm probability is the probability that the technique declares that Torah Codes are present in the case that there are none. This is the p-value of the experiment and we expect this to be very small, say less than 1/1000. And this is the value that both Torah code proponents and Torah code opponents are interested in.

The misdetect probability is the probability that the technique declares that Torah Codes are not present in the case that it is present. Since the Torah code opponents are not willing to entertain the hypothesis that Torah codes exist, since for them there can only be natural explanations, this probability is of no interest. However, this probability is of major interest to Torah code proponents because for a fixed false alarm rate, we desire that the misdetect rate be as small as possible. So when a Torah code opponent proposes an experiment with choices that would result in a large misdetect rate, the probability is increased that the experiment would provide evidence that Torah Codes do not exist when in fact they do exist. Such experiments would not be proper to run. For this reason not all experiments are equivalently good experiments.

If the situation is beginning to appear as technically complex, it is. It is technically complex and there is much research yet to be done to explore and characterize the operating characteristics of the different experimental choices. All this makes it more difficult to properly discuss the Torah Code controversy with a non-technical audience, particularly so given the logically incomplete arguments by the Torah Code opponents. For example, let us take at face value one of the arguments made in the McKay et. al. paper recently published in Statistical Science. In one of their experiments, they varied the measure of compactness and reported, in summary, that the particular measure employed in the original Witztum, Rips, and Rosenberg paper in fact worked the best. This they argue is evidence for WRR to have done an improper experiment, an experiment in which many measures were tried and they just found the measure that worked best with the Torah text and it was this best experiment that WRR reported. Logically, this argument is an argument that assumes the non-existence of Torah Codes and offers an explanation for the WRR experimental results that is consistent with this

assumption. There is another side to this argument. If we assume the existence of Torah Codes, then the fact that of the many compactness measures tried, the one used by WRR is the best just means that not all compactness measures are equally effective in detecting Torah Codes and that WRR happen to have been fortunate in settling early on to a good definition of compactness measure. What we have here is two self-consistent explanations. Logically, neither one can cancel the other.

Another variety of argument advanced by Prof. Simon goes like this. A proper experiment is one in which everything is specified in advance of the experiment and then the experiment is done. So if you do an experiment and if it were possible for there to be any wiggle room in the experiment, wiggle room where subjective choices in data gathering or in experimental protocol could have been made that could have made the experiment not an a priori experiment, "wiggles" that could have influenced the results to be falsely positive, then the experiment has no validity. The assumption here is that positive experimental results can only have occurred by natural means. And the only possible natural means is a non a priori experiment. Therefore, if there can be shown the possibility of "wiggle" room, then it can be assumed that the experimenter used it to obtain the positive results and in fact performed an invalid non a priori experiment. This explanation is self-consistent, but from it one cannot logically infer that a non a priori experiment was done in the case that there are possibilities other than natural means.

So if we examine the logic of the arguments of the Torah code proponents and the Torah code opponents, we will essentially find logical consistency. To rationally judge which argument is closer to the truth, first we have to separate the assumptions in each of the arguments. For the case of the Torah code proponents the assumptions basically come to the honesty of the experimenters doing a priori experiments and there having been no collusion among Witztum, Rips, Rosenberg, Havlin, Gans, and Inbal. For the case of the Torah code opponents, the assumption basically comes to the belief that only natural mechanisms are at play. Second, we have to see whether the picture painted by each side uses all the facts available or is selective. Without going into details, some of which are in this primer, it becomes clear that the picture painted by the Torah code proponents utilizes all the facts while there is some selective avoidance of certain facts by the Torah code opponents. So if one permits the possibility that in our reality there might be more than natural mechanisms at work, it is clear that the Torah code proponents have the stronger argument, because all the facts available are being used.

From the point of view of scientifically proving the existence of Torah Codes there is much work to be done. Indeed there are efforts underway now in gathering proper a priori data for future experiments as well as experiments to determine what is the best test statistic, alternative hypothesis to test against, and compactness measure to enable the detection of Torah Codes in any experiment. This work takes significant time and effort and I am sure that as experiments get completed and as some of the technical issues are better understood, there will be papers published describing what has been found out.

In terms of understanding what is known now, it is important to know how the Torah Code proponents answer the different general and technical arguments of the Torah Code opponents. In this primer, Mr. Gans thoroughly goes through the various technical arguments given by the Torah Code opponents. One by one he either gives alternative explanations or gives reasons for logical or statistical fault in the assumptions or conclusion of a Torah Code opponent argument. There is significant technical content to his counter arguments and the reading may not be so easy or quick. But it is essential to understand the nature of the arguments raised by the Torah Code opponents and it is essential to know that they are not logically complete. This means that they do not logically force a rational person to their stated conclusion as well as to know that they may only be considering that part of the experimental results supportive of their point of view. And it is important to know the strength of the positive experiments reported by the Torah Code proponents. The evidence offered is indeed enticing and suggestive that Torah Codes do exist. And I expect this evidence will get stronger and stronger as more and better experiments are done.

A Primer on the Torah Codes Controversy for Laymen

Harold J. Gans

Introduction

In August 1994 the peer reviewed journal "Statistical Science" published a paper entitled "Equidistant Letter Sequences in the Book of Genesis" by Doron Witztum, Eliyahu Rips, and Yoav Rosenberg. This paper reported the first scientific Torah codes experiment. As is well known, critics have challenged the validity and integrity of the scientific experiments that have shown the Torah codes phenomenon to be real. Some of the critics are technically qualified, and the challenges are often sophisticated. These attacks have been relentless, appearing in the media, and published widely in magazines and on the Internet. In addition, a peer-reviewed article by McKay, Bar-Natan, Bar Hillel, and Kalai (henceforth referred to as "MBBK"), "Solving the Bible Code Puzzle", purporting to prove that the Torah codes experiments are fatally flawed has been published in the May 1999 issue of Statistical Science (printed in September, 1999). We provide a concise summary of the major issues and their resolution in this paper. Note that one need not have any mathematical or scientific background, nor a background in Hebrew to understand what follows. One needs only to be able to follow logical reasoning and have patience.

Basics

Note: Some of the basics described below are themselves part of the controversy. These points will be discussed as appropriate later in this paper.

Doron Witztum and Professor Eliyahu Rips performed the first scientific Torah codes experiment in the 1980's. This experiment has five components, viz.:

1. The Hebrew text of the book of Genesis.
2. A list of famous Rabbinical personalities and their appellations (e.g., "Rambam" is an appellation of Rabbi Moshe Ben Maimon; "Hagaon MeVilna" and "HaGra" are appellations of Rabbi Eliyahu Ben Shlomo Zalman of Vilna, etc.).
3. A matching list of Hebrew dates of birth and death (month and day) paired with each personality (i.e., with each appellation of each personality).
4. A mathematical formula, which provides a measure of proximity between equidistant letter sequence (ELS, plural: ELS*s*) encodings of the appellations and ELS encodings of their corresponding dates of birth and death.
5. A mathematical technique that calculates the probability of obtaining the set of proximities obtained by the formula in number 4 above just "by chance". As required by one of the peer review referees, Professor Persi Diaconis, a probability against chance of 1/1,000 or smaller is considered a success. That is to say, suppose we hypothesize that the Torah codes do not exist. We then calculate a proximity measure. We now calculate the odds against the proximity measure obtained being at least as strong as it is. If these odds are 1,000 to 1 or greater, we have two possibilities: (a) An event with 1,000 to 1 odds just happened by chance, or (b) our initial hypothesis that the Torah codes do not exist must be wrong. Normal scientific procedure is to reject the possibility of an event with such strong odds happening by

chance. Such an event is so highly improbable that declaring it to be nothing more than chance verges on the absurd. In fact, the normal scientific standard for such a rejection is 20 to 1. Thus, we pick possibility (b), namely, Torah codes do exist and they cause the observed proximity measure to be as strong as observed. That is, we accept the statistical evidence as demonstrating that the phenomenon is real.

The Torah codes phenomenon, as discovered by Witztum and Rips, can be described as follows. Given an appellation of a famous rabbinical personality and his Hebrew date of birth or death, we search for each as an ELS in the Hebrew text of Genesis. If we find a minimum of one ELS of the appellation and a minimum of one for the date, then (some of) the ELS(s) of the appellation tend to be in closer proximity to (some of) the ELS(s) of the date than expected by chance.

Witztum and Rips prepared a list of 34 personalities, called "List 1", for their first experiment in the mid 1980's. These personalities were selected by a simple criterion, viz.: those that have at least 3 columns of text in their entry in the "Encyclopedia of Great Men in Israel; a Bibliographical Dictionary of Jewish Sages and Scholars from the 9th to the End of the 18th Century" edited by Margalioth and published in 1961. The list was then given to Professor S. Z. Havlin, of the Department of Bibliography and Librarianship at Bar Ilan University. Professor Havlin devised a set of historical / linguistic rules with which to determine the correct appellations associated with each personality in the list. In certain situations, his "professional judgement" was needed in addition to the rules. Professor Havlin thus produced a set of appellations, using his rules, which by his own testimony and that of Witztum and Rips, was produced with absolutely no feedback or coaching from Witztum or Rips. According to each of them, it was totally Havlin's work and completely independent of Witztum and Rips save for the latter providing the original list of personalities.

Witztum and Rips paired each of Havlin's appellations with each of three standard forms of the corresponding dates of birth and death (e.g. for the 13th of Adar the three forms are: rda gy, rdab gy, rda gyb) (for the 15th and 16th of the month there are 6 forms). This produced slightly under 300 pairs of appellations and dates. Using software developed by Yoav Rosenberg, they searched for each appellation and date in the list as an ELS code in the standard Koren edition of the book of Genesis. They next applied their proximity formula to those 157 pairs of ELSs found[1] (not all names or dates were found as ELSs in Genesis) and obtained results that appeared significant. (This appearance of significance was based on the erroneous perception that the proximity formula produced true probabilities.) However, at that time, a valid technique for assessing the probability (odds) of the result, the necessary 5th step, had not yet been developed.

The results were sent to Professor Persi Diaconis (through Professor David Kazhdan) who asked that a new list of personalities, List 2, be prepared from the same encyclopedia and the test redone on the new data. There were several possible reasons

[1] The formula was applied to all 157 pairs but produced usable results for the 152 pairs that had sufficient data size for part of the computation.

for this request: (a) If the result was valid, then a similar result would be expected on the new set. On the other hand, if the result on List 1 was really not significant but was just an anomaly, there would be no reason to expect to obtain another apparently significant result on List 2. (b)[2] There are many possible variations of the mathematical formula used to measure proximity. The possibility that many formulas were tried could not be ruled out. Clearly, the measure of significance depends strongly on how many formulas were tried; success in obtaining "interesting" results after trying many formulas is not nearly as significant as it would be had the same result been obtained by use of a single formula specified in advance (i.e., "a priori"). In fact, it is even plausible that the formula was purposefully constructed so as to give interesting-appearing results on List 1 (this would be "a posteriori", the opposite of "a priori"). Since the same formula would now be used on List 2 without any modification permitted, a success on List 2 could not possibly be attributed to any "fine-tuning" (or "wiggle room") of the formula to ensure a "success". (c) Although 3 common date forms were used in List 1, there are other (not so common) date forms. There are also several different versions of the Hebrew text of Genesis. Once again, since precisely the same date forms and Hebrew text of Genesis used in List 1 would be used for List 2, a success on List 2 could not possibly be attributed to the flexibility in choosing the date forms or the text. We shall henceforth refer to these properties as "list 1 – fixed". That is, any detail of the experiment on list 1 that was not changed before being applied to list 2 is "list 1 – fixed". A "list 1 – fixed" property or specification is thus not subject to "tuning" or manipulation on list 2 or any subsequent experiment.

List 2 was formed by selecting those personalities that have at least one and a half but less than three columns of text in their entry in the Margalioth encyclopedia. This produced a list of 32 personalities. Once again, Professor Havlin independently produced the corresponding appellations, and the final list consisted of slightly more than 160 appellation – date pairs found as ELS*s*. The proximity measure was applied to the els pairs found in exactly the same way as had been done with List 1. The results once again appeared significant.

In a September 5, 1990 letter addressed to Professor Robert Aumann, Professor Persi Diaconis, one of the referees for the peer review journal "Proceedings of the National Academy of Sciences", (PNAS), details a mathematical technique for measuring the probability of obtaining the set of proximities just "by chance" – that is, assuming that there is really no Torah code phenomenon. Some details that were not included in this letter (or were ambiguous) were included in a September 7, 1990 response from Professor Aumann. A copy of this letter was "given to Persi by hand in Sequioa hall, September 9, 1990, 2:50 PM. He looked it over and approved" according to a handwritten addendum on a copy of the letter, by Professor Aumann. Copies of these letters are included in appendix C. This was the necessary 5th component of the experiment. Professor Diaconis set the publication threshold for this probability at 1/1,000, well beyond the normal scientific standard of 1/20. That is, if the odds against getting such results (so many close proximities), given no Torah codes phenomenon was equal to or exceeded 1,000 to 1, the referees for PNAS would consider the result valid

[2] This was the only reason actually given by Diaconis.

enough so as to recommend publication. The test was applied (to List 2 only – by Diaconis' request) and the resulting probability obtained was 1/62,500, or more than 62 times better than the agreed upon threshold for publication. In spite of this, the referees decided not to recommend publication of the paper (for reasons related to the "scientific interest" of the result, rather than to the validity of the work).

The paper was subsequently submitted to the prestigious peer reviewed journal "Statistical Science". Several additional tests were required to ensure that the experiment did not produce apparently significant results spuriously. For example, the same test was done using a Hebrew translation of Tolstoy's "War and Peace" (the first 78,064 letters – the same as the length of Genesis). Had "significant" results been obtained here too (that is, a small probability, even if it were not as small as 1/1,000) this would imply that there is something very wrong with the methodology of the experiment. After all, no one claims that there are codes in "War and Peace"! This type of test is called a "control experiment". Other control experiments involved randomly mixing up the verses, or the words within verses, of the Hebrew text of Genesis. After successfully completing several control experiments (i.e., they produced random – looking results as they should), the paper was published by Statistical Science as "Equidistant Letter Sequences in the Book of Genesis" by Doron Witztum, Eliyahu Rips, and Yoav Rosenberg. This paper has come to be known as "WRR". It was August 1994.

Synopsis of the WRR experimental results

We present a few details of the experimental results obtained by WRR on list 1 and on list 2, as these details will be relevant later. An individual proximity measure, $c(w, w')$ was calculated for each pair of words w and w′ in each list. $c(w, w')$ ranges from near 0 (actually, 1/125) to 1, and the smaller it is, the stronger the proximity between the ELSs of the paired words. Thus, 1/125 is the strongest individual proximity measure, and 1 is the weakest. An overall proximity measure can then be calculated for the entire list. For list 1, two such overall proximity measures were calculated. The first of these (called "P1") is given as a "sigma" value[3]. The larger this number, the stronger the overall proximity. This measure is based on how many of the individual proximity measures, $c(w, w')$, are less than or equal to 0.2 (recall, individual proximity measures are stronger when the $c(w, w')$ values are smaller). Since we will refer to this quantity (here, 0.2) several times in this paper, we shall call it the "P1 bunching threshold". That is, P1 is a function of the number of $c(w, w')$ that bunch below this threshold. The second overall proximity measure, called P2, is based on all the c values, and the smaller it is, the stronger it is (opposite to that of P1). The same two overall proximity measures were used for list 2. In addition, each was modified slightly to produce two more overall

[3] In statistics, this is related to the normal distribution. However, in the context of these experiments that relationship does not hold. The underlying cause of the breakdown in this relationship, namely dependencies among the individual proximity measures, was not perceived by WRR when the experiment was done on list 1. Hence, they thought that the normal probability associated with P1 was a true probability. Occasionally, this "probability" is still given as P1 instead of the sigma value, although with the understanding that it is not a true probability. Similarly, WRR also thought that P2 was a true probability at that time.

proximity measures; the 3rd (P3) being similar to the 1st, and the 4th (P4) being similar to the 2nd. A description of how P3 and P4 are defined will be given later.

Proximity measures

	Name	Description	Remark
Individual measure	c(w, w′)	One c value per word pair	Smaller values are stronger
1st Overall measure	P1	Based on c values ≤ 0.2	Bigger values are stronger
2nd Overall measure	P2	Based on all c values	Smaller values are stronger
3rd Overall measure	P3	Similar to P1	Bigger values are stronger
4th Overall measure	P4	Similar to P2	Smaller values are stronger

For list 2, each of the four overall proximity measures is converted into a true probability, the smallest (most significant) being 4/1,000,000. We will describe this process later in the paper. Finally, the four probabilities are converted into one overall probability by simply multiplying the smallest of the four by 4. Thus, the overall probability for list 2 is 4 × (4/1,000,000) = 16/1,000,000 = 1/62,500. The following table gives the values obtained for list 1 and for list 2. Note that small numbers are given in scientific notation: 1.29E-9 is the same as 0.00000000129, 1.15E-9 is the same as 0.00000000115, and 7.20E-9 is the same as 0.00000000720. The symbol "σ" indicates a "sigma" value.

Experimental results

	List 1	**List 2**
1st overall proximity	6.61σ	6.15σ
2nd overall proximity	1.29E-9	1.15E-9
3rd overall proximity	---	5.52σ
4th overall proximity	---	7.20E-9
1st probability	---	453/1,000,000
2nd probability	---	5/1,000,000
3rd probability	---	570/1,000,000
4th probability	---	4/1,000,000
overall probability	---	16/1,000,000 = 1/62,500

Non scientific challenges

The validity of each of the 5 components of the WRR experiment has been challenged. In addition, questions have been raised concerning a tradition for the existence of ELS codes in the Torah, and the appropriateness of teaching codes, particularly as part of an outreach program. Concerning these non scientific issues, suffice it to say that several leading *Gedolim* (Torah sages) have given their written approbations stating that there is a tradition of Torah codes, as well as the appropriateness of teaching Torah codes for outreach. In a 1989 *psak din* (judgement), Rav Shlomo Fisher stated "It is a mitzvah to teach codes". Although some critics claim that in private conversations since then, Rav Fisher has indicated a retraction of that position, there is nothing in writing to lend credence to such a claim. On the contrary, in 1997, after discussing the issues at length with a number of critics, as well as with Doron Witztum, Rav Fisher wrote another letter of support and further indicated that he,

together with Doron Witztum, had gone "a number of times in front of Rav [Shlomo Zalman] Auerbach, Zt'l, and he also backed this without reservation". In 1998 more letters of approbation were obtained from Rav Shmuel Deutch, Rav Shlomo Fisher, Rav Shmuel Auerbach (son of Rav Shlomo Zalman, Zt'l), Rav Shlomo Wolbe, and Rav Mattisyahu Salomon. In particular, Rav Wolbe states, "It is known that a way exists to discover hints and matters from the Torah by reading letters at equidistant intervals. This method is found in the commentary of Rabeinu Bachya on the Torah and in the works of Rav Moshe Cordovero". In 1999, Rav Shmuel Kamenetsky wrote a strong letter of support for Torah codes subsequent to his having had an extended conversation with some of the critics followed by a discussion with Harold Gans. Rabbi Moshe Heinemann has also stated that teaching Torah codes is a *kiddush Hashem*. He has further indicated that anyone who doubts this or challenges it should feel free to call him directly and discuss it. Anyone concerned with these issues should read the letters of these *Gedolim*. An English translation of the full text of these letters can be found in appendix B. The remainder of this paper will deal with the technical challenges to WRR.

Challenges to the text of Genesis

The Hebrew text of Genesis used for the WRR experiment (and indeed for all subsequent experiments unless specifically noted otherwise) is the Koren edition. This is the text used by Jewish communities all over the world (except the Yemenite Jews). (Incidentally, the experiment also produces statistically significant results on the Yemenite edition of Genesis, albeit somewhat weaker.) The fact that the worldwide standard text was used rules out any concern that the choice of the text was not a priori, i.e., that many different manuscripts or versions were used and Koren simply worked best. Note too that the text used was "list 1 – fixed" as explained previously, and therefore cannot be an issue for the experiment on list 2. In fact, the challenge comes from a different direction. There are "biblical scholars" who claim that the standard Koren edition of the Torah has a plethora of incorrect or missing letters. Of course, these views are supported by interpretations of statements in the Talmud or elsewhere. Now, since a single letter missed or added to the text will destroy an ELS code, even if there had been ELS codes in the Torah at some point in time, none could possibly have survived to be found in the Torah available to us today. Therefore, claims concerning the discovery of ELS codes in the Torah must be wrong.

Resolution

It must first be noted that the question posed is related to the issue of Torah codes, but cannot be a challenge to them. The evidence for Torah codes is obtained from analysis of the text of Genesis available to us today. No claim is made concerning how the codes got there or what the text looked like in the past. Hence, at most, the "challenge" simply asks the question, "How could there be Torah codes in the text if there is evidence that the text has been corrupted?" This question is an important one but it does not challenge the existence of the codes. One can surely find many possible answers to this question. We will show, however, that the basic premise of the question is likely to be incorrect.

Let us first consider the effect that a dropped or added letter in the text has on an ELS in the text. It is true that a single added or dropped letter will destroy an ELS – provided that the error is somewhere between the first and last letters of that ELS. It does so by making the distance between the two letters flanking the error unequal to the remaining equidistances (+1 for an added letter, -1 for a deleted letter). Clearly, an error that is outside the span of an ELS can have no effect on that ELS. Furthermore, the smaller the distances between the letters of an ELS, the less likely it is that an error will occur in its span – and the WRR experiment assigns more weight to ELS*s* of relatively small span as opposed to those of larger span. Furthermore, several ELS*s* in the text usually represent each appellation or date. What this means is that a few errors are unlikely to destroy enough ELS*s* to render the entire WRR experiment no longer significant. Experimentation bears this out. On the other hand, hundreds of errors in the text of Genesis surely would destroy any evidence of codes. We can now rephrase the critic's argument more accurately, as follows: (1) A large number of errors in the text of Genesis would destroy any Torah codes phenomenon. (2) There are authoritative opinions that there are thousands of letters incorrect or missing in the Torah. Therefore, codes cannot exist in the Torah we have today. All we need to do now is to reverse the argument: WRR provides very strong evidence that there are codes in the Torah. Hence, either (1) or (2) above must be incorrect. Since we know (1) to be true, the problem lies with (2). That is, there cannot be large numbers of errors in the text of Genesis. The "authorities" quoted by the critics must be wrong. Since the opinion held by these "authorities" is by no means the only authoritative opinion on that subject, this is certainly a viable and reasonable alternative. MBBK admit that there is a variance of opinion on the accuracy of Genesis: "The amount of variation that has already occurred during the many preceding centuries since Genesis was written is a matter of scholarly speculation..." (MBBK, page 165). In fact, some authorities believe that there is no more than 1 error in the entire Torah[4]. We certainly cannot resolve this issue here, but the existence of Torah codes implies that those who believe that there are very few errors in Genesis are closer to the truth than those who believe that there are many errors. In any case, one can certainly not disprove Torah codes by assuming a position that is the subject of a heated debate by the experts in that field. For more details on the historical evidence for textual integrity, please refer to the excellent article by Rabbi Dovid Lichtman, included as appendix A.

It is interesting to note that this particular challenge concerning the accuracy of the text of Genesis has implications to another question posed by the critics. They point out that ELS*s* for some appellations do not appear at all in Genesis. Furthermore, even when they do, they are not always in close proximity to an ELS of the matching date. Sometimes they are even in closer proximity to an ELS of an incorrect date! How valid could the claim of a close proximity Torah code then be?

The question posed reveals a misunderstanding of the WRR experiment; once one understands the experiment the apparent difficulty vanishes. WRR makes no claim that all appellations or dates are found as ELS*s* in Genesis. Nor is there a claim that every

[4] The questionable letter is in Deuteronomy chapter 23, not in Genesis, and is thus irrelevant to the WRR experiment.

appellation and matching date will have ELS*s* in close proximity even if they do have ELS*s* in Genesis. WRR simply makes the following claim: If one searches for ELS*s* for each appellation in the list, and for the corresponding dates, and measures the proximity between the ELS*s* found for the appellations, and the ELS*s* found for the matching dates, then too many of them are in close proximity to each other to ascribe to "chance". This claim is not at all in conflict with the critic's observation that not all ELS*s* are found, and not all are in close proximity. Consider the following example. Given any English text that contains the words "umbrella" and "rain", we would surely expect these words to appear in close proximity more often than expected for unrelated words. We might even be able to measure this effect statistically. But no one would claim that every appearance of "rain" must be close to an appearance of "umbrella"! Still, one might pose the critic's question as a philosophical question rather than as a challenge. There are several possible answers. Perhaps the answer most likely to be correct is that we have not yet perfected our techniques and understanding of the codes phenomenon to the point where we can be assured of detecting everything that is encoded. The research into the Torah codes phenomenon is in its infancy. It is interesting to note, however, that one possible answer is that some ELS*s* may have been lost due to a small number of errors in the text. Thus, the critics' question is answered by their own challenge!

Hidden failures

We will now discuss an important concept known as "hidden failures". Suppose one performs an experiment, and one obtains a probability of 1/10,000. This means that the result obtained is unlikely to have happened by chance. Specifically, it means that the expectation is that one would have had to perform 10,000 different experiments before such a result might happen by chance alone. If one performs an experiment and obtains such a significance level, one is justified in concluding that it did not happen by chance. Rather, something caused this unusual result to happen, e.g., the existence of a code. Suppose, however, one did 10,000 experiments and one of them yielded a significance level (probability) of 1/10,000. Such a result is quite expected and we cannot conclude that anything (e.g., a code) more than chance is operating here.

Consider a simple example. Suppose you think of a random number between 1 and 100 and challenge a friend to guess the number. If he succeeds on the first try, then it is startling because by chance alone the probability of him doing so is 1/100. Suppose he guesses the number on the second try. This is still interesting but not quite as startling as guessing it on the first try. We can actually calculate the significance (mathematically, the "expectation") by simply multiplying the probability of success with one guess by the number of guesses. In this case it is $2 \times (1/100) = 1/50$. Suppose he takes 50 different guesses? Then his probability of success is $50 \times (1/100) = ½$ and this is not significant at all! If he takes 100 different guesses he is sure to guess the right number; his probability of success is $100 \times (1/100) = 1$.

We see from this discussion that the number of experiments performed is just as important as the probability of success calculated because the true measure of success is the product of these two numbers. Thus, if one claims to have done an experiment and

obtained a probability (of success given that exactly one experiment was performed) of 1/10,000 but has actually done, say, another 499 experiments in secret, then the experimental results have been misrepresented. He should have reported that he did 500 experiments before obtaining the 1/10,000 result, in which case the true significance of what he has found could be calculated as $500 \times (1/10{,}000) = 1/20$ – still significant by the accepted scientific standard, but not nearly as significant as reported. Such an experiment is a fraud and is said to contain "hidden failures". In this case there were 499 hidden failures.

An example of where the charge of hidden failures was made concerns a tape of a lecture given by Professor Rips in the mid 1980's, (around the time when the WRR experiment on list 1 was first performed). The existence of this tape was publicized by McKay in an Internet posting dated December 31, 1997 and entitled "New Historical Evidence Available". In this paper McKay says, " Several aspects are particularly disturbing, as they don't appear to fit in with the 'official history' of the experiment". McKay then proceeds to explain, "The lecture describes an experiment with 20 rabbis, defined by having an entry in Margalioth of at least 6 columns. We find that difficult to reconcile with the statement, repeated many times, that the first test ever done was with 34 rabbis having at least 3 columns". In other words, it would appear from this tape that at least one hidden experiment was performed and never documented. If there is one hidden failure, one may assume that there are many more besides the one that has been fortuitously uncovered.

Resolution

Note first that this challenge is irrelevant to the experiment on list 2, and the published WRR result is for list 2 only. Nevertheless, we can resolve this issue for list 1 as well. Indeed, if one sets the criterion for personality selection at 6 columns of text or more in Margalioth, one produces 20 personalities. Thus, there is evidence here that another experiment, different, but closely related to the list 1 experiment, was at least contemplated at that point in time. This, by itself, is not unusual. It would be hard to imagine that WRR performed the first experiment that came to mind. The critical question is whether or not the experiment (and perhaps others as well) was actually performed. If we examine the words of Rips on the tape, it certainly appears that such an experiment was performed. For the quotes that follow, we use McKay's posted English transcript of the tape of the original Russian lecture. Rips says, "If an article took up three or more pages, we chose it" (page 13). Note that Rips actually says "3 pages", not "6 columns" as McKay quotes him. They amount to the same, but it is easier to see someone confusing 3 pages with 3 columns than someone confusing 6 columns with 3 columns. Undoubtedly, McKay made this little change without realizing that it made any difference. Rips continues, "We had an objective list which turned out to contain some 19 – 20 names" (page 13). So the list was definitely made. Rips proceeds to talk about using appellations and dates and concludes, "Thus we have obtained a list containing some 150 queries" (page 14). But wait! There are only 102 pairs (queries) for this list of personalities. There are 152 pairs – for list 1. This is the list produced by using entries with 3 columns or more, not 3 pages or more! We thus have a contradiction in the tape itself. Rips talks about 19 to 20 personalities, but the number of

pairs obtained corresponds to the 34 personalities of list 1, not the list of 20 personalities. Rips then describes the result of the experiment, "Therefore, on the basis of all this, we have a significant...7 sigmas...or let us say 6 sigmas...never mind...6 sigmas is quite sufficient to state with absolute confidence that what we have here is a real phenomenon..." (page 15). Now the score quoted here, between 6 and 7 sigmas, is P1, the first of the two overall statistics calculated for list 1.The sigma value calculated for list 1 of the WRR experiment is 6.61, while the sigma value for Rips' 20 personalities is 5.74. Thus, not only does the number of pairs quoted by Rips correspond to list 1 of the WRR experiment, so does the result. The evidence thus suggests that although an experiment with 19 – 20 personalities was certainly considered, the experiment actually performed was on the 152 pairs of list 1. It is understandable that Rips might have confused "3 columns" and "3 pages" (but not 6 columns!) while giving a lecture. Incidentally, a P1 score of 5.74 sigma is extremely significant[5]; it is hard to imagine why anyone would change the list in any way having obtained such a strong result. Perhaps it is because of these considerations that MBBK are content to simply state, "However, an early lecture of Rips (1985) described an experiment with a particular subset of '19 or 20' rabbis. Be that as it may...." (MBBK, page 153) without pursuing the implications!

One more point can be made. As indicated earlier, the overall proximity measure of 5.74 sigma is calculated in part by counting how many individual proximity measures, $c(w, w')$, are less than or equal to 0.2, the "P1 bunching threshold". The value 0.2 appears to be arbitrarily set. Had WRR done just a bit of experimentation, they would have discovered that setting this value to 0.5 rather than 0.2 changes the overall proximity measure from 5.74 sigma to 6.19 sigma. In terms of sigma values, 6.19 is considerably stronger[6] than 5.74. The fact that WRR "lost" this opportunity for a more significant result will figure prominently when we discuss the MBBK paper and "tuning". As we shall see there, changing the P1 bunching threshold from 0.2 to 0.5 makes the P1 measure for the entire list 1 7,000 times stronger. Nevertheless, WRR did not avail themselves of this obvious opportunity to "tune" their experiment to increase the apparent significance of their result.

Additional experiments

Before we address other challenges to the WRR experiment, we will discuss some additional Torah codes experiments that were successful. We will only describe four of these experiments, although there are a good deal more, because three have direct implications to the validity of WRR, and the fourth is so remarkable that it would be remiss not to include it.

[5] At the time, WRR erroneously thought that this score was normally distributed. As such, it would represent a significance level of 4.73E-9.

[6] As indicated in an earlier footnote, these sigma values are not normally distributed. If they were, 6.19 sigma would be 16 times stronger than 5.74 sigma. Recall that at the time, WRR thought that the sigma values of P1 were normally distributed.

The cities experiment

We will first describe the "cities experiment" of Gans, and its history. In the late 1980's, Harold Gans, then a Senior Cryptologic Mathematician with the National Security Agency, US Department of Defense, was told about the Great Rabbis Experiment. Being skeptical, he requested that Witztum and Rips provide him with the Book of Genesis on a computer disk so that he could duplicate the experiment. A few months later the data was provided. Gans did not immediately rerun the experiment; he reasoned that the data would never have been provided if the experiment were fraudulent. However, in 1990 Eric Coopersmith, then head of Aish HaTorah in North America, requested that he attempt to duplicate the Great Rabbis Experiment. Gans did so, using his own programs and following the specifications of the experiment in a preprint of the WRR paper. He then conceived of a new experiment: to use the same names and appellations as in WRR's list 1 and list 2 combined, but pair them with the names of the cities of birth and death, as opposed to the dates of birth and death as in WRR. He asked Zvi Inbal, a new acquaintance and a lecturer for Arachim in Israel to provide the list of cities for the new experiment. Inbal obliged, providing Gans with the list, along with an outline of the methodology used to construct the list. The database for the list of cities was the same encyclopedia used by WRR, in addition to the Encyclopedia Hebraica. The text of Genesis, the mathematical formula for proximity, and the method of calculating statistical significance used were precisely the same as in WRR. (Actually, Gans first used a minor modification of WRR's proximity measure, but later abandoned it in favor of WRR's formula.) Gans completed the cities experiment in 1990 and documented his results in a preprint entitled "Coincidence of Equidistant Letter Sequence Pairs in the Book of Genesis". The results were even more significant than that obtained by WRR: 6/1,000,000, or about 1/166,000. The paper was submitted for publication in Statistical Science but was rejected because it was not considered "of interest to the broad audience of Statistical Science" (Letter from the editor of Statistical Science to Gans, July 25, 1995). The editor suggested that a paper "whose focus was a review of the literature on probabilistic numeralogic calculations in Biblical texts" would be of more interest!

In 1997, critics suggested to Gans that the Inbal list of cities had been contrived to ensure an apparently significant result. They pointed out that many names in the list were spelled differently than in either Encyclopedia and that there were apparent inconsistencies in the spellings. Gans took these criticisms seriously. He announced publicly that he had started an investigation of the validity of the Inbal list and would not take an official position on its validity until the investigation was complete. (Gans continued lecturing on codes during the investigative period, but always pointed out that the experiment was under investigation. He also pointed out that nothing had yet been found to suggest that the experiment was flawed.)

The first step in the investigation was to obtain a detailed explanation of the methodology of obtaining the city names. Upon request, Inbal provided Gans with a detailed explanation of the rules used to generate the list. These rules, now referred to as the "Inbal protocol", form a complete algorithm which can be applied in a purely mechanical way to form the list from the two encyclopedias. The Inbal protocol is quite

complex, enabling it to produce linguistically and historically correct data from encyclopedias that are inconsistent in spelling the names of foreign cities transliterated into Hebrew. Another important issue (among several) is that many of the cities had specifically Jewish names. Since WRR used Hebrew as opposed to secular dates, consistency demanded that Hebrew names be used for the cities even where the encyclopedias used secular names. Inbal also provided Gans with a detailed explanation of how each name/spelling in the list was obtained using the protocol. This list contained a handful of corrections to the original list.

The task before Gans was twofold: first, to verify that the Inbal protocol, with all its complexity, was not contrived to ensure an experimental "success", but rather was designed solely to ensure linguistic and historical accuracy. Secondly, to verify that each name/spelling on the list was obtained by a purely mechanical application of the protocol. These tasks took Gans two years to complete. First, he extracted all the rules comprising the protocol, and posed these as questions to rabbaim, dayanim (Jewish Judges), roshei kollel (deans of advanced Torah study institutions), and experts in writing gittin (Jewish divorce documents, which must include the name of the venue of the divorce properly spelled in Hebrew or Aramaic) in the US, Israel, and England. He only queried those who had no previous contact with any of the details of any Torah codes experiments. For each question posed, he received one of two answers: either they did not know, or their answer agreed with the Protocol. There was not one instance in which anyone felt that the Inbal rule was wrong! There was also no rule in the protocol that was not verified by some of the experts queried. In addition, no one felt that the protocol was incomplete and should have additional rules[7], with one exception. One expert suggested that a single alternate form should also be tried: the addition of the digraph "q q" (an acronym for *k'hal kadosh*, meaning "holy community") as a prefix to each name[8]. This was tried and failed. The conclusion of this part of the investigation was clear: the Inbal protocol was designed solely to ensure historical and linguistic accuracy. It was certainly not contrived simply to produce an apparent experimental success.

The next task was to verify the Inbal list. For this, Gans obtained the assistance of Nachum Bombach. Together, they applied the Inbal protocol and produced a list that they compared to that of Inbal. They noted several differences. In two instances Inbal used other historical sources rather than the two encyclopedias because in these cases the protocol produced erroneous data. In one case it was due to a typographical error in one of the encyclopedias. In the other case, it was due to two cities having very similar names. Gans and Bombach decided to use the erroneous data rather than violate the protocol. Thus, although there are a few known errors in the list, it is produced solely by mechanical application of the Inbal protocol. There are no exceptions. The final list produced by Gans and Bombach differs from the original Inbal list in only a handful of

[7] This is not to say that there aren't some exceptions to the rules. However, there are no additional rules that would cover all the exceptions and not produce incorrect spellings as well.

[8] This abbreviation is not pronounced; it is exclusively a written form. This is excluded by Havlin's rules. Inbal simply followed his precedent. None of the experts consulted were familiar with any of the details of the WRR experiment, and certainly not with Havlin's rules.

places. A new experiment was run on this list and produced the same level of statistical significance as had been obtained on the original Inbal list: 6/1,000,000. The investigation was successfully completed, and was announced publicly in May 1999 at the International Torah Codes Society conference in Jerusalem. Gans is currently documenting this work with all its details so that anyone who wishes can easily verify both the Inbal protocol and the list.

A replication of the famous rabbis experiment

Doron Witztum performed two other successful experiments that we shall describe. The first is entitled "A Replication of the Second Sample of Famous Rabbinical Personalities". The idea that underlies this experiment was suggested by Alex Lubotsky, a critic, in an article entitled "Unraveling the Code" in the Israeli newspaper "*Ha' aretz*", September 3, 1997. In this experiment, the personalities and dates are exactly the same as in WRR's list 2. Instead of using the appellations of WRR, however, each personality was referred to as "*ben name"* (son of *name*), where "*name*" is the name of the personality's father. The father's names were obtained from the same encyclopedia used for the data in WRR. The spelling rules used were exactly the same as specified in writing before the experiment was performed by Havlin and used for WRR. (In fact, the spellings are identical to the Margalioth entries with only two exceptions). No other appellations were used. Once again, the text of Genesis and all the mathematical components of the experiment were precisely the same as in WRR. The significance level obtained was 1/23,800 – an unqualified success. Witztum also tried using appellations of the form "*ben rabbi name*", but the result was not significant. It is also worth noting that the same two experiments were tried on list 1. The *"ben rabbi name"* experiment had a significance level of 0.0344. This difference in significance level and form between list 1 and list 2 is an unexplained curiosity, but does not detract from the unambiguous success obtained with list 2. Recall that WRR reported its success on list 2, not list 1.

Personalities of Genesis

The second additional experiment performed by Witztum that we will discuss is entitled "Personalities of Genesis and their Dates of Birth". Witztum found that dates of birth for some personalities in Genesis are given explicitly in several Talmudic period sources. Using the CD-ROM of the Bar Ilan Responsa Database, he searched all the sources to find the one source that gave the largest number. (Recall that data size is often a critical element in trying to detect statistical significance.) The *Yalkut Shimoni* provided the largest sample: 13 personalities (Adam, Yitzchok, and 11 of Yaakov's sons). Upon examining a critical edition of the *Yalkut Shimoni*, Witztum found that all 12 sons of Yaakov were included, making the final data size 14. An experiment similar to WRR, measuring proximity between personalities and dates of birth as encoded in Genesis was performed. The date forms used were exactly the same as in WRR. The personality names were spelled exactly as found in Genesis. The assessment of statistical significance was also done exactly as in WRR. However, the formula for measuring proximity was slightly different from the WRR technique in one point: the skips of the ELS*s* of the personalities were restricted to +1 and –1, i.e., strings of letters

in the text itself, both forward and backward[9]. The dates were searched for in exactly the same way as in WRR. This approach was not new, as it had been used in an earlier experiment of Witztum and Rips known as the "Nations experiment" and recorded in a preprint in 1995 (we shall describe this experiment shortly). Two versions of the experiment were performed; one produced a significance level of 1/1,960 while the other was 1/21,740. Once again, an unambiguously significant result was obtained. Even if one claimed that the original WRR proximity measure must have been tried as well, this would only multiply the result by 2, giving a significance level of 1/10,870.

A second version of this experiment, "The Tribes of Israel" (also known as "The sons of Yaakov"), was also performed by Rips. Even accounting for the possibility that separate experiments were performed on data taken from each known source, rather than from a single source (*Yalkut Shimoni* for Witztum, *Midrash Tadshe* for Rips), the significance level was still 1/28,570! One would have to claim that there were experiments on at least another 200 data sets taken from different (unknown) sources before one could discount the result as insignificant because of the number of experiments tried to achieve a success. No one has even suggested that anywhere near that many additional sources of this information as an explicit list (each of which must be different, of course) exist.

It is interesting to contrast the approach of Witztum and that of Rips in dealing with the potential challenge that there were hidden failures in this experiment. There are several different sources for the information, as well as different options in choosing how to spell the names. Witztum's approach is to make a single choice of all the parameters, and explain how he made his choice and why it is logical that only that choice was made. Rips' approach is to tally all the sources and arguable choices that are possible and show that even if all of them were tried, the results are still very significant.

There have been other successful scientific Torah codes experiments performed by other, independent researchers, including Nachum Bombach, Alex Rottenberg, and Leib Schwartzman. Thus, other independent researchers have reproduced the Torah codes phenomenon. We have chosen to describe the three experiments above because they have logical implications to the validity of the WRR experiment. We now describe a remarkable result obtained by an associate of Witztum.

The nations prefix experiment

In 1995, Witztum, Rips, and Rosenberg (WRR II) prepared a paper entitled "Equidistant Letter Sequences in the Book of Genesis II: The Relation to the Text". This experiment has come to be known as the "Nations experiment". Most of the details of this experiment, and the critic's challenges to it are not relevant to the result that we are concerned with here, so we will only describe those details that are needed. Genesis, chapter 10, lists 70 descendants of Noah's three sons. These 70 descendants were the progenitors of the 70 biblical nations that constitute all of humanity. 68 of the 70 names are distinct. For the Nations experiment, each nation name is paired with a defining

[9] This choice is a natural one since the names of the personalities of Genesis appear in the text of Genesis itself.

characteristic of a nation, used as a prefix to that nation's name. Four prefixes were used for what is known as the "regular" component of the experiment: \u (nation), {ra (country), tpc (language), and btk (script). Thus, for example, for }unk (Canaan) the four pairs would be: (}unk, }unk \u) (Canaan, nation of Canaan), (}unk, }unk {ra) (Canaan, country of Canaan), (}unk, }unk tpc) (Canaan, language of Canaan), and (}unk, }unk btk) (Canaan, script of Canaan). WRR II provide reasons for selecting these four defining characteristics based on a commentary on the Book of Job written by the Vilna Gaon, the Targum Yonatan's names for the nation/country as well as the plural forms. They also provide justifications for the Hebrew words used for these characteristics. MBBK (page 162) dispute the a priori selection of these prefixes and argue that many characteristics and words representing these characteristics were tested, and high scoring ones were then chosen as prefixes. An a posteriori "story" was then concocted to "justify" the given selection. Unlike WRR, where the proximity is measured between ELS*s* of the pairs of words, here the proximity is measured between appearances of the first word of each pair in the text itself, either forward or backward (i.e., skip distances +1 and –1)[10], and ELS*s* of the second word. A few minor modifications were made in the proximity measure to accommodate this change. The final probability reported for this "regular" component of the experiment as described here was 70/1,000,000,000 for P1 and 5,667/1,000,000,000 for P2.

On March 11, 1998, Bar Natan, McKay, and Sternberg (BMS) posted a paper on the Internet entitled "On the Witztum – Rips – Rosenberg Sample of Nations" (draft). In this paper, BMS construct a "counterfeit" Nations experiment to demonstrate how WRR II could have done the same. Several responses to this paper were posted by Witztum in which he explains all of the details of how the Nations experiment was performed, and refutes the arguments of BMS. We are concerned with an experiment performed by an associate of Witztum, and reported in Witztum's August 29, 1999 paper "The Nations Sample" (Part II). BMS construct their experiment by starting with a list of 136 possible prefixes (BMS, Table 5), including the four used by WRR II. They next rank the 136 prefixes using the P2 proximity measure and select a high scoring set of four. They also construct a story to "justify" their choices. BMS obtain an apparent significance level of 5/100,000,000 using a Hebrew translation of "War and Peace"[11] (the same used as a control by WRR).

BMS point out that there may be mathematical problems with the technique used to measure the probabilities by WRR II (BMS, Appendix; Notes on the Metric). Consequently, Witztum replaced the original measurement scheme with one that does not have any of these potential problems. This technique, called RPWL (Random Permutation of Word Letters), was devised by Witztum in 1994 for "Header Samples" and was described in a letter to Professor Robert Aumann on 22 February 1994. It was used again for WRR3 in 1996. The RPWL technique was posted on Witztum's Web site

[10] Since the names appear in the text of Genesis itself.

[11] Actually, the control text used by BMS is "War and Peace" with the part of Genesis that lists the 70 nations replacing the corresponding section of "War and Peace". This ensures that the names are found in the text as they are in Genesis. The necessity of doing this is not contentious (provided the decision to do this was a priori).

in 1997. The paper describing the application of RPWL to the WRR list 2 is entitled "The 'Famous Rabbis' Sample: A New Measurement" and is dated May 10, 1998. Using this measure, the overall significance of the "regular" component of the Nations experiment is 1.4E-7 (i.e., 14/100,000,000) while the BMS "significance level" is 1.22E-4 (i.e., 122/1,000,000). BMS's "significance level" is meaningless since it was obtained by "tuning" (cheating). BMS claim that WRR II did the same.

The experiment performed by Witztum's associate uses the 136 prefixes created by BMS along with the accurate RPWL measurement scheme. Each one of the prefixes is ranked with P1 and P2, both for "War and Peace" as used by BMS, and for Genesis. The top scoring 4 prefixes are then selected for each. For "War and Peace" the score obtained for the combination of the best 4 prefixes is 6.16E-6, while for Genesis it was 4.0E-10. Neither of these is a true significance level (probability) since the four prefixes for each were not chosen a priori, but rather they were chosen to optimize the proximity measure. Yet, they differ by almost 4 orders of magnitude. The real issue is this: given the optimization procedure used, what is the true probability of obtaining a score of 4.0E-10 (or stronger)? Using this procedure, is a score of 6.16E-6 within the range of random expectation? It will be recalled that the individual proximity measures, c (w, w′), range from near 0 to 1 in value. One million simulations of this experiment were performed by simply replacing c (w, w′) with a computer generated pseudo-random number with values ranging uniformly from near 0 to 1. The ranking of the "War and Peace" score of 6.16E-6 among these million simulations was 353,949. Thus, the probability of obtaining such a small score (or smaller) by taking the best four out of 136 prefixes is 0.35. This is well within the range of random expectation. The ranking of the score 4.0E-10 obtained on Genesis among the million simulations was 420. Thus, the probability of getting such a small score (or smaller) on Genesis is 0.00042. This probability easily passes Diaconis' threshold of 1/1,000. This result is remarkable because there are no stories to justify, no choices of prefixes to justify, and the list of 136 prefixes was provided as an apriori list by BMS. This result also suggests that there may be more encoded in Genesis concerning the 70 nations than WRR II suspected since the four best prefixes of the 136 in Genesis include two not used by WRR II.

Challenges to the date forms

WRR used three date forms. For example, for the 13th of Adar the three forms are: rda gy, rda gyb, and rdab gy. No one argues that these three forms are incorrect, or unusual, but there are other forms that exist. In particular, the form rdab gyb is not uncommon. In addition, there are several other date forms that can be found in use, e.g., rdal gy, rdal gyb, rda lc gy, rda lc gyb but are rare. There are also variations in the names of the months (e.g., Cheshvan vs. Marcheshvan), etc. It would appear that the issue of hidden failures would be relevant here. WRR might have performed many hidden experiments using various combinations of the date forms, and then reported the combination that worked best.

Resolution:

The issue of hidden failures does not apply here at all. Since the date forms were completely specified for list 1, there was no freedom (sometimes called "wiggle room")

to change them for list 2. That is, the date forms were "list 1 – fixed". Consequently, although theoretically there might have been wiggle room utilized in choosing the date forms for list 1 (which is one of the reasons that experimental results were not reported for list 1 in WRR), this was logically impossible for list 2.

It is worth noting that the form rdab gyb does well on list 1 – much better than the form rda gy (MBBK, pg. 156). If wiggle room had been utilized for list 1 then it should have included the form rdab gyb. We conclude that it is logically impossible for any wiggle room with the date forms to have been taken advantage of for list 2, and there is good evidence that it did not take place for list 1.

One last point: the question of date forms does not apply at all to Gans' cities experiment. As for Witztum's two experiments listed above, the date forms used there are identical to the ones used in WRR. As with list 2, there was no wiggle room at all.

Challenges to the proximity formula

The proximity formula is an essential component of any Torah codes experiment. This formula accomplishes the following: Given a list of pairs of words (e.g., names and dates of death), along with the positions in the text where these words are found as ELS*s*, it produces a small set of numbers. Each of these numbers (P1, P2, etc.) is a slightly different measure of the overall proximity between ELS*s* of the paired words of the list. Each of the overall proximities of the list is, in turn, composed of the individual proximities ("c (w, w′) values") found between ELS*s* of the words making up each pair. For WRR there were 4 overall proximities calculated; for the cities experiment and the two additional experiments of Witztum described above, only (the last) 2 of the 4 were used.

Each overall proximity obtained must next be input to the process which calculates the probability of obtaining that measure by chance, i.e., assuming that there is no Torah codes phenomenon. The 4 (or 2) probabilities thus obtained may then be combined in a standard way to yield a single probability measure for the entire experiment. For WRR, the four overall proximity measures are (1) 6.15 sigma, (2) 1.15E-9, (3) 5.52 sigma, and (4) 7.20E-9. The probabilities of obtaining each of these proximity measures at random are (1) 453/1,000,000, (2) 5/1,000,000, (3) 570/1,000,000, and (4) 4/1,000,000. (The reason for presenting these probabilities as fractions over 1,000,000 will be made clear later in this paper.) These 4 probabilities are then combined to yield 16/1,000,000, which is equal to 1/62,500.

The proximity formula used is complex. This implies that there are many parts of it that could be changed to produce different proximity measures. This opens the door to hidden failures. Perhaps a large number of variations of the formula were tried, and only the one that worked best was publicized.

Resolution:

The resolution here is identical to the resolution of the challenge to the date forms since the proximity formula was "list 1 – fixed". The same holds true for the additional

experiments (except for a possible factor of 2 in the "personalities" experiment, as described above).

Challenges to the process that calculates the probability

In the previous section of this paper, we indicated that the last major step in any Torah codes experiment is to transform the overall proximity measures into probabilities. These probabilities are then combined in a standard, non-controversial way, to give the final probability for the entire experiment. The probability thus calculated is the probability that a single experiment would produce the given overall proximity measures (or better) if there were no codes in the Torah. For example, given the first overall proximity measure of 6.15 sigma for the WRR experiment, the probability of a single experiment producing such a measure (or better) is 453/1,000,000. This means that one would expect to get a proximity measure that strong or stronger approximately 453 times for every 1,000,000 experiments performed, assuming pure chance, i.e., no codes.

The process of calculating the probability associated with an overall proximity measure is simple and straightforward. The overall proximity measure is comprised of all the individual proximity measures between personality appellations and dates of birth and death. Consider the effect of randomly mismatching the personalities and the dates (i.e., randomly permuting the dates vs. the personalities). The appellations are the same, the dates are the same, the text of Genesis is the same, and the proximity measure is the same, but the information is wrong. For example, suppose there were 5 personalities. For simplicity, we assume that each has one appellation and one date. Then we might represent the data as follows:

Personality(1)	date(1)	proximity(1,1)
Personality(2)	date(2)	proximity(2,2)
Personality(3)	date(3)	proximity(3,3)
Personality(4)	date(4)	proximity(4,4)
Personality(5)	date(5)	proximity(5,5)
		Overall proximity

A "mismatched" or "permuted" experiment (one in which the dates have been randomly permuted with respect to the personalities) might look like this:

Personality(1)	date(5)	proximity(1,5)
Personality(2)	date(3)	proximity(2,3)
Personality(3)	date(1)	proximity(3,1)
Personality(4)	date(2)	proximity(4,2)
Personality(5)	date(4)	proximity(5,4)
		Overall mismatched proximity

There are many ways of mismatching the personalities and dates, particularly for large numbers of personalities as in WRR. If we perform, say 99 permutations then we have 1

real overall proximity and 99 overall proximities for permuted experiments, or 100 all together. If there are no Torah codes, then there is no meaning to any of these proximities, and the real proximity measure is expected to be similar to the ones for the permuted experiments. It follows that the probability of the real overall proximity being the strongest of the 100 is 1/100 since there are 100 places and only 1 is best. Similarly, if one performs 999 random permutations with corresponding overall proximity calculations, and the real one is best, the probability is 1/1,000. If it is second best, the probability is 2/1,000 or 1/500, and so on. In this way, we can directly compute an estimate of the probability of obtaining a given real overall proximity measure by actually performing many permuted experiments and simply counting how many overall proximities for permuted experiments are better than the real one. As the number of permuted experiments is increased, the accuracy of this estimate is also increased. This process is often called a "randomization". For WRR, Diaconis requested that 999,999 permuted experiments be performed (letter to Prof. Robert Aumann, Sept 5, 1990). Together with the real experiment, this makes 1,000,000 experiments. A probability of 453/1,000,000 for the first overall proximity measure, P1, means that the real measure ranked 453 out of 1,000,000. Similarly, a probability of 4/1,000,000 for the 4th overall proximity measure, P4, means that only 3 overall proximity measures (of the same type, i.e., P4) on permuted experiments out of 999,999 ranked better than P4. (The 1/62,500 = 16/1,000,000 = 4 × 4/1,000,000 is the probability of obtaining 4 probabilities, the smallest of which is 4/1,000,000.)

This approach is so intuitive and simple, it is hard to see where it might have a problem. Indeed, the challenge does not claim that the technique is totally incorrect, but rather that because of certain circumstances in the WRR experiment, the value that it produces may be inaccurate. One of the circumstances that might effect the accuracy is the fact that some dates and names appear more than once in the list (producing "dependencies"). A second circumstance that might effect the accuracy is the fact that the real experiment happens to have more pairs than about 98% of the permuted experiments and this may give the real experiment an advantage that has nothing to do with the presence of codes.

A further challenge to the calculation of the probabilities is the charge that Prof. Diaconis did not suggest the method; rather, it was Witztum and Rips who suggested it through Prof. Aumann. Specifically, MBBK say, "However, unnoticed by Diaconis, WRR performed the different permutation test described in section 2" (MBBK, page 153). Thus, the process for calculating the probability might have been designed by Witztum and Rips to ensure a "success" on list 2 (and presumably on list 1 as well).

Resolution

It is important to note that MBBK do not feel that this challenge is sufficient to reject the WRR experiment. We quote directly from page 154 of their paper: "Serious as these problems might be, we cannot establish that they constitute an adequate 'explanation' of WRR's result". Note: not only can they not establish that this "problem" constitutes a fatal flaw, they do not even claim that the "problems" are serious, only that they "might" be serious! There is good reason for their lack of a forceful challenge. They

cannot prove that these "problems" are serious because there is strong evidence that these "problems" are not serious at all. First, we note that if the process were fatally flawed, it is impossible to explain why it should work on so many different experiments (WRR, Gans' cities, the additional Witztum experiments, etc.) but it does not work on any of the control experiments. If the process produces spurious results, why does it favor real experiments as opposed to permuted experiments? After all, whatever repeated words there are in the names and dates (i.e., dependencies) are equally present for all the experiments, real and permuted! Thus, the real experiment can have no advantage over the permuted experiments as a result of dependencies.

Let us now consider the "problem" of data size. The critics point out that the real WRR experiment has more pairs than 98% of the permuted experiments[12]. The critics claim that "the effect of this extremeness is hard to pin down..." (MBBK, page 154). In fact, the effect is not at all hard to pin down because it is easily eliminated. Their hypothesis is that experiments with larger data size (more pairs) have a statistical advantage over experiments with smaller data sizes. This statistical advantage could result in the correctly matched WRR experiment scoring better than (98% of) the permuted experiments even if there are no codes. Note first that no reason is given as to why this should be so. To test if this hypothesis is true, all we need do is restrict the permuted experiments to the 2% that have equal or larger data size than the real experiment. If the rank of the real experiment still yields a probability less than 1/1,000 then we conclude that "the effect of this extremeness" is not responsible for the success of the experiment, and the "problem" is no problem at all.

McKay first raised this issue on February 22, 1997 in a paper posted on the Internet and entitled "Equidistant Letter Sequences in Genesis – A Report (DISCLAIMER PRELIMINARY DRAFT ONLY)". In this paper McKay performs an experiment similar to the one suggested in our analysis above. He restricts the permutations to those that have exactly the same size as the real experiment. He reported a P2 rank of 127/1,000,000 (page 4) and notes that this is "24 times worse" than the WRR result. What he failed to point out in that paper is that 127/1,000,000 = 1/7,874, which is still more than 7 times more significant than Diaconis' success threshold of 1/1,000. (In a later draft, this value is no longer quoted and Mckay notes that there may have been some errors in the data used for the first draft.) Furthermore, if one uses P4 – the measure that did best in WRR – not P2 as McKay did (why didn't he use P4?), and one requires that all permuted data sizes be greater than or equal to the unpermuted data size, one obtains a probability of 1/1,000,000. If one does 100,000,000 permutations (to obtain more accuracy) then the probability is 28/100,000,000, as opposed to a probability of 72/100,0000,000 if one does not constrain the data sizes. We conclude that there is evidence that extremeness in data size, first discovered by McKay, adversely affects the accuracy of the WRR results. When the extremeness is removed, the results are more significant! Apparently, MBBK felt that it was important to raise the question concerning extremeness in data size, but that it was irrelevant to mention the solution which McKay himself had proposed over two years earlier!

[12] This is a theoretical value. In reality it is somewhat less than 98% for list 2.

For the cities experiment the data sizes of the real and mismatched experiments are much closer than they are in WRR. If one applies the fix described above so that the "problem" does not apply, one obtains the same significance level as before: 6/1,000,000.

There is a further test that was applied to corroborate the accuracy of the process for computing the probability. A set of 30 control experiments (for the cities experiment) was performed by randomly mixing up (permuting) the letters within each word in the list. The permuted version of each word replaced the original word wherever it appeared in the original list. Thus, the number of repeated words (and therefore, dependencies) and the data sizes were maintained exactly. The 30 results were then analyzed to see if they conformed to random expectation. Even a slight bias could be significant if it were present in all or even most of the 30 results. The final probability measuring the significance of any deviation from random expectation in the 30 results was 0.4. This is a totally insignificant probability and is strong evidence that these so-called "problems" are not problems at all. On random data, uniformly random results are produced[13]. Significant results are obtained only for real data. We now understand why even the critics did not claim that these problems are "serious", but only that they "might" be serious. It is also clear why the critics admitted, "we cannot establish that they constitute an adequate 'explanation' of WRR's results". One cannot establish as true that which can be empirically demonstrated to be untrue!

Another problem that has been raised (MBBK, page 171) is the following. Some appellations and dates are not found as ELS*s* in Genesis. Suppose all appellations (or all dates) for a particular personality are not found. Then it would seem that that personality could be removed from the list since it does not contribute any individual proximity measures. WRR, however, kept these entries, allowing the dates (or appellations, if no dates were found) to be paired with other appellations (dates) in the permuted experiments. According to MBBK, this introduces "noise" in the final probabilities.

In this case, MBBK may be correct, and as they show in their paper, removing these personalities from the list improves the final probability slightly. This is another opportunity for an improved experiment that WRR missed. We shall deal with the implications of this observation later.

There is another refutation of MBBK's thesis that various biases skew the accuracy of the randomization process that calculates the probability from the proximity measures. In a May 1998 paper posted on the Internet and entitled "The 'Famous Rabbis' Sample: A New Measurement", Witztum uses a measure first introduced in 1994 to calculate the probabilities in a way that avoids all the biases noted by MBBK. Part of the process involves partitioning the list of appellation -– date pairs into three sets. Each set uses only one of the three date forms, thus avoiding each appellation appearing multiple times corresponding to the different date forms. There were also several other changes including a randomization based on random permutations of the letters comprising each date rather than permutations of the dates versus the personalities (the RPWL technique

[13] In mathematical terms, on non-causal data, the process produces results that are uniformly distributed over the unit interval.

discussed earlier). The final probabilities obtained for each set were: Set 1: 0.000000626, Set 2: 0.0258, and Set 3: 0.012. Thus, if the biases noted by MBBK had any effect on the results it was a detrimental one; it did not enhance the result. It is interesting to note how MBBK deal with this randomization technique – publicized a year before MBBK posted their paper on the internet -- a technique that proves that the WRR result is not due to biases in the methodology. MBBK do not mention it at all! Contrast this with MBBK's assertion that "nothing we have chosen to omit tells a story contrary to the story here" (MBBK, page 152).

The origin of the permutation test

We now address the charge that the permutation test used to obtain the probabilities was not suggested by Diaconis, but rather was invented by Witztum and Rips and suggested to Aumann. **Note that in any case, this method was "list 2 – fixed" and hence could not be manipulated for any of the additional experiments.** However, we will demonstrate that the charge is untrue. Appendix C contains copies of several letters between Diaconis and Aumann that conclusively demonstrate that Diaconis suggested the permutation test. The reader is urged to read these letters. In the September 5, 1990 letter to Aumann, Diaconis states that they "are in agreement about the appropriate testing procedure for the paper by Rips et al." He then goes on to describe the permutation test. Some of the details are clarified in the September 7, 1990 response from Aumann to Diaconis. In a paper entitled "The origin of the permutation test" by McKay and Kalai (MK) and posted on McKay's Web site shortly after the appearance of the MBBK paper, they state, "For each Rabbi, there were a list of names and dates. The permutation test works like this: calculate some measure of the average closeness of the names and dates. Randomly permute the dates..." Here is the first problem. Nowhere does Diaconis ever suggest an "average". The word is used by MK but not by Diaconis or Aumann. Diaconis does talk about an "additive" test, but Aumann clarifies this in his September 7 letter as follows: "incidentally, 'bunching' or 'twenty percent' might be a more suggestive name for the test you call 'additive' ". "Bunching" and "twenty percent" are clearly referring to the proximity measure P1 which measures the "bunching" of $c(w, w')$ that are less than 20%, or 0.2. MB would have us believe that there are two different tests referred to in these letters, a "Type A" test suggested by Witztum and Rips through Aumann, and a "Type D" test suggested by Diaconis. MB define these tests as follows:

> "Type A: Take all 300 or so name – date pairs (from all the rabbis), then combine those 300 distances into an overall measure.
> Type D: For each rabbi, combine the distances of his name – date pairs into a single number, then combine those 32 numbers into an overall measure."

Nowhere does Diaconis explicitly call for a "Type D" measure. Diaconis would not call for such a test because "Type A" and "Type D" refer to the proximity measure, not a probability. The proximity measures P1 and P2 were defined by Witztum and Rips and thought, at the time, to be probabilities. Diaconis objected to the use of P1 and P2 as probabilities because it involved an implicit and unproved assumption that the individual proximity measures, $c(w, w')$, were independent. Therefore Diaconis and Aumann discussed how to measure the significance of P1 and P2 (as well as P3 and P4). The permutation test does this. However, if the individual $c(w, w')$ for all appellations and

dates are combined into a single measure for each personality, then the process for calculating the probability is not being applied to P1 and P2, but rather to a new function of the c(w, w′)'s. That is, a new proximity measure has been created and the significance of the new measure is then computed. The success or failure of such an experiment is irrelevant to the original question posed: the significance level of Witztum and Rips' P1 and P2. Note too, that P1 and P2 were fixed on list 1 and no modifications were permitted for its application to list 2. This was the reason Diaconis requested the test on list 2!

Let us examine the May 7, 1990 letter written by Diaconis to Aumann. The second sentence of the letter is "I have four points to make on the test proposed by Witztum, Rips, and Rosenberg". Thus, Diaconis is not suggesting anything new; he is commenting on what WRR had proposed. In paragraph 2 (the second of the four points) Diaconis says, "as I understand it[14], there is a fixed set of matched pairs (x_1, y_1) (x_2, y_2),...,(x_n, y_n) and a statistic $T = d(x_1, y_1) + ... + d(x_n, y_n)$". Thus, Diaconis does refer to a sum, but it is not his suggestion. This is what he understood concerning the proximity measure that Witztum and Rips had already applied to list 1. Diaconis is certainly not saying "as I understand it" with reference to his own suggestion! It would seem that at the time Diaconis thought that there was only one individual proximity measure for each personality, probably because he did not realize that there was more than one appellation and date per rabbi. Note, however, that this was his understanding at the time, not his suggestion. Nowhere does he define "$d(x_i, y_i)$" as MK would have us believe – Diaconis assumed that Witztum and Rips had already done so and accepted their definition of proximity, whatever it was.

We see that Diaconis was mistaken concerning the definition of the proximity measures, thinking that the individual proximity measures (his "$d(x_i, y_i)$") were summed. He does not suggest that an average be taken to combine the $c(w, w')$ into $d(x_i, y_i)$ for each personality. This would be redefining the proximity measure – something that neither Aumann nor Diaconis wanted since any change would result in the WRR hypothesis remaining untested. There is another reason Diaconis would not make such a suggestion – it makes no mathematical sense unless the measure is being tuned to guarantee failure. This is because when small values (significant) are averaged with large values (insignificant), the large values dominate the sum and the significance is lost. Consider the following example. Suppose we have two probabilities. One is $p_1 = 1/1{,}000{,}000$ while the other is $p_2 = ½$. We wish to calculate a single measure of the significance of obtaining both p_1 and p_2. Using the same technique as used in WRR, we obtain $p = 2 \times (1/1{,}000{,}000) = 1/500{,}000$. On the other hand, if we take the average of the two, we get 0.2500005, which is insignificant. It follows that Diaconis would never have made such a suggestion unless he was a novice in statistics or purposely tuning the measure to fail – hardly a reasonable thesis. Diaconis does refer (NOT suggest) an "additive" test and a "multiplicative" test. As we have shown, the "additive" test corresponds to the P1 measure while the "multiplicative" test corresponds to the P2 measure.

[14] This author's emphasis.

How do MB maintain a thesis that Diaconis and Aumann were talking about two different ways of calculating the probability when Diaconis and Aumann never say a word about such a disagreement, and actually say several times that they were in agreement? One can understand Diaconis not paying attention to the proximity measure. It was designed by Witztum and Rips and was not to be altered. But as for measuring the probability, this was the central issue that Diaconis was addressing! Here is MB's explanation of this paradox: "Neither person acknowledged the difference between them. It was almost as though neither person was reading the letters written by the other" (MB, page 6). It is important to note that the WRR paper was written before the calculation of the probability on list 2 was done. Of course, the results of the computation were not included, but were represented by question marks. This paper was given to Diaconis by Aumann, and Diaconis approved it (see the appendix for a copy of critical pages from this paper). In a cover letter for the WRR paper, dated Dec 6, 1991 (see appendix), Aumann states "needless to say, the test itself was not changed in any way; it is precisely the one to which we agreed in the summer of 1990." The WRR paper contains an exact and detailed description of the proximity formulas P1, P2, P3, and P4, as well as the permutation test. It was only after Diaconis approved the paper that the calculation of the probability was done. We may assume that Diaconis, one of the referees, noticed that the proximity formula was not what he thought it was back in May of 1990, but it made no difference. The proximity formula was never the subject of discussion between Aumann and Diaconis; the calculation of the probability was. Surely Diaconis would not have approved an unauthorized change in his probability calculation. (There is even another letter of agreement between Diaconis and Aumann dated August 28, 1992 concerning further research and again defining the probability calculation as in WRR.) Finally, when the WRR paper was rejected by Diaconis and the other referees, it was rejected because it was not considered appropriate for PNAS; the results were not considered to be of scientific interest. It was not rejected because the experiment performed did not meet Diaconis' approval or specifications.

It is interesting to note that although Diaconis never mentions what the definition of "d (x_i, y_i)" is for each personality, McKay does offer a possible definition. On August 9, 1998, in a paper entitled "Revisiting the Permutation Test", McKay defines this quantity as "the average of the logarithms of the defined c(w, w′) values". He then writes: "Use of the logarithms gives much stronger prominence to the word pairs with small distances and in my opinion meets the objection of Rips that ordinary mean "averages out" the alleged ELS phenomenon". McKay then presents the results of this experiment: "For the WRR data, this method gives very respectable scores of 125/ million for the first list and 8/ million for the second." Thus, McKay himself provided a possible definition for the quantity that Diaconis never defined and obtained very significant results! Somehow, MBBK "forgot" to mention these results in their paper in spite of their earnest statement that "Nothing we have chosen to omit tells a story contrary to the story here".

It is noteworthy that it is possible to combine all the individual proximity measures for each personality and obtain a single true probability for each personality. This was first done with list 2 by Nachum Bombach using a technique called "BST"

(Best Star Team) developed by Professor Robert Haralick. These individual probabilities can then be combined using a standard statistical formula (Fisher's Statistic) to obtain the probability associated with the entire list. The probability obtained this way is 0.00000218, very close to the original WRR result. On the other hand, if this technique is applied to MBBK's "War and Peace" "experiment", the result is only 0.00225 – it doesn't even pass the 1/1,000 threshold! Thus, this technique appears to distinguish between the real and the counterfeit.

We conclude that either (a) MB have misinterpreted the historical record, or (b):

(i) Diaconis said "as I understand it" in reference to his own suggestions, and

(ii) Diaconis suggested things that make no mathematical sense or were specifically designed to fail, and

(iii) Both Aumann and Diaconis respond to each other's letters over an extended period of time without appearing to read the mail they receive. They say that they are in agreement on the central issue – calculating the probability of the proximity measures P1 through P4 – when in fact they do not even know what the other is talking about! and

(iv) Diaconis approved the WRR paper before the probability calculations were performed without noticing that the calculations described in detail were not what he specified!

We will let the reader judge between these two alternatives. Finally, recall that this issue is irrelevant to the cities experiment and all the other additional experiments since the calculation of the probability was fixed once the experiment on list 2 was documented.

Challenges to the appellations

We now come to the most serious charge: that the appellations used were selected specifically to produce an apparently significant result for list 1 as well as for list 2. We have already seen that virtually all the parameters of the WRR experiment were fixed on list 1 and thus could not be manipulated to produce a "success" on list 2. There is, however, one major component of the experiment that was new for list 2: the appellations. Thus, logically, if any component of the experiment on list 2 was manipulated to affect an apparent success, it has to be the appellations. Furthermore, in order to prove that such manipulation is possible and practical, MBBK produced another set of "appellations", "similar" to those of WRR's list 2, that "succeeds" very well on a Hebrew translation of Tolstoy's "War and Peace" (in fact, the same used as a control experiment by WRR) (MBBK, page 157). Since MBBK were able to select "appellations" to make their experiment look successful, WRR presumably could have done the same to make their experiment look successful on Genesis.

Resolution

Before we resolve this issue, it is worth making a few observations. Witztum and Rips have declared publicly that they did not provide any input at all to the selection of appellations. They say that they did no more than provide Havlin with the list of personalities. Havlin also has "certified explicitly that he had prepared the lists on his

own" (MBBK, page 156). It follows that subconscious manipulation of the appellations to affect a "success" is not possible in this case. Witztum and Rips had no control over the choice of appellations, and Havlin had no way to assess which appellations would contribute to the success or failure of the WRR experiment. If a fraud were perpetrated here, it must have been a conscious collusion between Witztum, Rips, and Havlin. This is not impossible, but it is unlikely that a world class academician (Rips) and a renowned rabbi and scholar (Havlin) would risk their professional and personal reputations by perpetrating a hoax which was bound to be revealed eventually. As we shall see, if there is a conspiracy here the number of people necessarily involved in it will stretch the credulity of any reasonable person.

It will be recalled that in the above section entitled "non scientific challenges", several *Gedolim* (leading rabbis and sages) wrote letters of support for WRR[15]. In a testimony co-authored by Rav Shmuel Deutch and Rav Shlomo Fisher, they state, "We checked the rules according to which Professor Havlin formulated his list of names and titles of Torah leaders, and we found that it was commensurate with both professional standards and common sense. The list is in keeping with the principles. We found that all[16] the opponents' individual claims concerning deviations from the principles are false, and are a testimony to their glaring ignorance and unfamiliarity with the subject. In light of the above, we hereby affirm that the work of Rav Doron Witztum, Professor Eliyahu Rips, and Rav Professor Shlomo Zalman Havlin does not contain an iota of fraud or deception and the claims of their opponents are a reprehensible libel". In other words, Havlin's rules and lists, as well as his "professional judgement" were checked by two independent experts, *Gedolim*, and certified to be correct. They specifically state that there were no deviations from the rules. Yet, MBBK claim that list 2 does not follow Havlin's rules completely and use this as justification for breaking Havlin's rules to form their list for "War and Peace" (MBBK, page 157). MBBK had to break Havlin's rules because the wiggle room supposedly introduced by those instances where Havlin used "professional judgement" was insufficient to allow tuning the experiment to the extent required. The only way MBBK could make their experiment "work" on "War and Peace" was to deviate from Havlin's rules. This implies that any wiggle room resulting from Havlin's use of "professional judgement" was also insufficient to tune the WRR experiment to the extent required. It follows that MBBK's demonstration that one can manipulate the appellations to fit an arbitrary text falls apart! Their "demonstration" could only be accomplished by breaking Havlin's rules, whereas WRR did not break the rules. In addition, MBBK claim that since Havlin's rules were only made public after the WRR experiments were complete (MBBK, page 157), they were retrofitted to the lists after the fact. Note that these two claims are contradictory: either the rules existed before the lists were formed or the rules were retrofitted to the lists – but not both! If the former is true, then MBBK's "War and Peace" list would have to follow pre-existing rules. If the later is true, MBBK has to retrofit rules to their list. But they have done neither! Given the *Gedolim*'s statement that "...all the opponents' claims concerning deviation from the principles are false...", it is clear that each and every appellation in list 2 (the list that they checked) is consistent with Havlin's rules. There are 102 appellations in list 2. The

[15] An English translation of the full text of all of these letters can be found in appendix B.
[16] This author's emphasis.

possibility of constructing rules that are "commensurate with both professional standards and common sense" and that will also fit "all" of 102 appellations, presumably selected to make an experiment look successful, is extremely remote. The ability to do something like that would be a wonder in itself! The critics have never demonstrated that such a thing is possible or even plausible. Unless one includes these *Gedolim* in the Witztum – Rips – Havlin "conspiracy", the charge of fraudulent appellation selection has been totally refuted.

In an article entitled "My Cities Experiment – Analysis and comments" posted on the Internet (http://wopr.com/biblecodes), Dr. Barry Simon says, "Mr. Gans suggests the right resolution 'would be to accept the challenge issued by Doron Witztum in an article in Galileo some time ago. He suggested that an independent linguistic expert whom everyone concerned agrees is impartial and qualified should be asked to provide a new list of Rabbis and appellations. Such an experiment would test the original Rabbis experiment directly. To date, no one has accepted Mr. Witztum's challenge.' But Mr. Gans fails to note a critical aspect of Mr. Witztum's proposal which is 'to allow an independent authority to prepare a new list of names and appellations for the 32 personalities on the second list, *using Prof. Havlin's guidelines*.' Namely, after the wiggle room has been favorably frozen by the rules (only stated nine years after the original experiment), there is a proposal to ask an outside expert to come in constrained by these rules. This is hardly a test of the rabbis experiment – it is a charade." It is clear from this statement that the critics acknowledge that if an impartial and independent expert were to reproduce the list of appellations strictly following Havlin's rules, the experiment would succeed! Otherwise, the challenge would have been eagerly accepted. Dr. Simon claims that the "wiggle room has been favorably frozen by the rules" which were constructed to fit the list. Hence, deviation from Havlin's rules is not the issue. The critics are claiming that it is possible to retrofit rules to appellations. This implies that their counterfeit experiment on "War and Peace" completely misses the point. The critics would have to retrofit rules such that (a) Every one of the appellations used in their "War and Peace" "experiment" follows the rules, and (b) the rules can be confirmed as being "commensurate with both professional standards and common sense" by independent and impartial authorities. Until the critics can meet the challenge of accomplishing precisely that which they claim WRR accomplished, their "War and Peace" demonstration is vacuous. This conclusion follows logically from the critic's own statements.

It is of interest to note that the challenge concerning the appellations can also be refuted on logical grounds without being an expert in Hebrew or bibliography. Gans' cities experiment uses precisely the same appellations as in lists 1 and 2, and produces very significant results. If the appellations were selected on the basis of their being in close proximity to the dates, why are they also in close proximity to the cities? Recall that Gans' idea of using city names came years after lists 1 and 2 appeared in preprints. It follows logically that the success of the cities experiment implies that the appellations in lists 1 and 2 were not selected on the basis of close proximity to the dates, i.e., the appellations were chosen honestly, and both the WRR experiment and the cities experiment are valid successes. The obvious retort is that the city names were selected to be in close proximity to the appellations. In this scenario, the conspiracy has grown to

include Witztum, Rips, Havlin, Rav Deutch, Rav Fisher, and Zvi Inbal, who provided the list of cities to Gans, as well as Gans and Bombach who verified the authenticity and accuracy of the protocol and list. Recall, however, that every single item in the city list is produced by the Inbal protocol without exception. This is easily verified, and, in fact, is not challenged by the critics in their paper. The protocol itself was made public years ago by Inbal and it, too, has not been challenged. No "counterfeit" cities experiment in "War and Peace" (or anywhere else) has ever been successfully performed by the critics. In order to show that such an experiment could be faked, MBBK would have to produce their own linguistically correct protocol and follow it mechanically to produce an apparent success in "War and Peace". This has never been done. It is instructive to see just what MBBK do say about the cities experiment. On page 163 they say, "The only other significant claim for a positive result is the preprint by Gans (1995), which analyzes data given to him by an associate of Witztum. It was later withdrawn (Gans, 1998), but Gans recently announced a new edition which we have not seen. The original edition raises our concerns regarding the objectivity of the city data, as many choices were available". In other words, there is no direct challenge to the protocol or the list. Rather, the author of the experiment has himself "withdrawn" the experiment, so it surely must be invalid.

Had Gans truly withdrawn his experiment? MBBK refer to a statement made by Gans in 1998. It is instructive to see what Gans actually said: "This unwillingness to speculate on an outcome of an investigation while it is still ongoing has prompted some people to interpret that as evidence that I am no longer convinced that the Torah codes phenomenon, as detailed in WRR, is a real phenomenon or that I no longer believe that the conclusions drawn from my original cities experiment are correct. Let me then state in absolute terms that this is not true. To date, I have not uncovered a single fact or even a hint that the list of cities that I was provided was manipulated in an attempt to make the results of the experiment appear significant when, in fact, they are not significant. I have not uncovered a single fact that causes me to doubt that the conclusions drawn from the original cities experiment were accurate." (Gans, March 23, 1998). Does this sound like a withdrawn experiment? This sordid affair can thus be succinctly summarized as follows. The critics charged that the cities list was not honestly produced by Inbal. Gans responded by announcing that he would investigate their charges, and the critics then claimed that Gans had withdrawn his experiment – in spite of Gans' public declaration to the contrary. Compare MBBK's version of Gans' statement with their statement that "Nothing we have chosen to omit tells a story contrary to the story here" (MBBK, page 152). Furthermore, on May 11, 1999, a full month before the release of the MBBK paper on the Internet, Gans announced at the International Torah Codes Conference in Jerusalem, in the presence of one of the authors (Bar Hillel) that he had completed his investigation and found that the protocol and list were accurate (except for a handful of errors that were corrected), and the results very significant – 6/1,000,000. There was no challenge from Bar Hillel. Thus, Gans confirmed the Inbal protocol and list (except for a handful of errors); he did not announce a "new edition" that MBBK did not see. The protocol and list have been available to the public for years[17]. MBBK claim that "many

[17] It is of interest to note that the quote from MBBK given above is from the second version of that paper. The original version read, "At the time of this writing it has been withdrawn by its author 'to conduct a

choices were available", but do not list any! The most telling flaw in MBBK's case against the cities experiment is that they claim many choices were available, presumably to allow the experiment to be tuned to "succeed". Yet MBBK have never succeeded in using these "choices" to tune a counterfeit cities experiment in "War and Peace". MBBK produce no challenge to the protocol or the list, but base their rejection of the experiment on the claim that the author himself has withdrawn the experiment, along with some vague and unsubstantiated notion of "many choices" being available. MBBK have been challenged publicly on several occasions to prove that the cities experiment could have been produced by taking advantage of wiggle room, by using the purported wiggle room to produce a counterfeit cities experiment in "War and Peace". They have so far refused the challenge.

The experiment "A Replication of the Second Sample of Famous Rabbinical Personalities" by Witztum described in the earlier section "Additional experiments" provides a second line of independent proof that the WRR experiment is not a hoax. In this case, the appellations were replaced by the simplest appellations possible: "*ben name*" ("son of *name*") where "*name*" is the name of the personalities' father as obtained from the Margalioth encyclopedia. No other appellations were used. The spelling rules used were exactly the same as specified in writing by Havlin and used for WRR. (In fact, the spellings are identical to the Margalioth entries with only two exceptions.) Every other component of the experiment is exactly the same as in WRR. Thus, there is no wiggle room in any of the components of the experiment. The probability obtained was 1/23,800.

The experiment "Personalities of Genesis and their Dates of Birth" by Witztum (and the version "Tribes of Israel" by Rips) provides a third line of independent proof that names and dates of birth are encoded in Genesis. The names are spelled exactly as found in the book of Genesis, and all other components are exactly the same as in WRR (except for one component taken from the "Nations" experiment of Witztum and Rips. As described earlier, this introduces at most a factor of 2 in the final results). The probability obtained here is at least 1/10,870 (including the factor of 2). It follows that none of the challenges raised applies to this experiment. The date forms, the proximity formula (with the factor of 2), the calculation of the probability, and the text of Genesis are all fixed from WRR and there are no appellations.

How does MBBK deal with these two experiments and their direct implication to the validity of WRR? They just ignore them! There is no mention of either of these experiments anywhere in their paper, even though both experiments were made public months before the MBBK paper appeared. In fact, MBBK, after discussing several older experiments, state, "The only other significant claim for a positive result is the preprint of Gans (1995)..." (MBBK, page 163). Given three independent experiments with no wiggle room that prove that the WRR experiment could not possibly be a hoax, MBBK deal with

---

thorough investigation of every aspect of the city selections and spellings' (Gans, 1998)". Thus, MBBK did not even mention Gans' announcement of a "new edition" in the original version of their paper. It was changed as a result of a letter of protest sent by Gans to the editor of Statistical Science. Note, too, MBBK's selective quote from Gans' statement, totally removed from context.

them by claiming that one was withdrawn by its author, and ignoring the existence of the other two! Nothing could be a stronger testimonial to their inability to find a flaw in any of these experiments. The absence of a "counterfeit" to the cities experiment, reveals that the critics have not demonstrated the possibility of selecting city names and spellings to insure an experimental "success" while strictly adhering to a protocol. We conclude that Rav Deutch and Rav Fisher were absolutely correct when they declared that the rules and lists of Havlin "do not contain an iota of fraud or deception".

Let us briefly summarize what we have determined. Since the text of Genesis, the date forms and the proximity formula are "list 1 – fixed", none could have been manipulated in any way to affect the outcome of the experiment on list 2 or any of the additional experiments described (cities, personalities in Genesis, replication of the list 2 sample, the nations prefix experiment). Since Professor Diaconis specified the method of calculating the probability, it too was not manipulated to affect the result. The method of calculating the probability is also "list 2 – fixed" and could not have been manipulated to affect the results of the additional experiments. As for the appellations, the following points are relevant:

1. The appellations were "list 1 and list 2 – fixed" and so could not have been manipulated for the cities experiment.
2. The cities names/spellings were produced by a strict application of the Inbal protocol without exception. The protocol and list are not challenged by MBBK. The critics have never demonstrated the feasibility of producing a counterfeit cities experiment.
3. The "replication" experiment uses no appellations at all – just the true names of the fathers of the personalities with the prefix "*ben*". This experiment directly confirms the results on list 2. This experiment is not challenged by MBBK.
4. The "personalities in Genesis" experiment uses no appellations – just the spellings of the names as found in the text of Genesis itself. This experiment provides a success similar in form to WRR. It is not challenged by MBBK.
5. Several *Gedolim* have verified both Havlin's rules as well as the appellations used in list 2, derived through the application of those rules.
6. The claim that the critics have "done the same" in "War and Peace" is false. They have admitted explicitly that their list is not consistent with Havlin's rules. Hence, their experiment bears only a superficial resemblance to the WRR experiment. Furthermore, the critics claim that the rules were retrofitted to the list (which is why they will not agree to an independent trusted expert forming a new list using Havlin's rules). Since they did not retrofit any rules to the list used in their "War and Peace" experiment, their experiment bears only a superficial resemblance to their view of the WRR experiment and proves nothing.
7. In the nations prefix experiment, the appellations and dates are replaced by a list provided by McKay et al. and hence were not subject to any manipulation by WRR.