BS"D Draft. 30 Tishri 5761 (29 Oct. '00)

MBBK'S "STUDY OF VARIATIONS"

DORON WITZTUM AND YOSEF BEREMEZ

INTRODUCTION

The famous experiment of WRR concerning the hidden code in *Genesis*, which was published in *Statistical Science* [1], is the subject of the critical paper of MBBK (McKay, Bar-Natan, Bar-Hillel & Kalai): "Solving the Bible Code Puzzle," published in the same journal [2]. MBBK try to prove that the result of WRR's experiment is not valid. In the introduction to their paper they write:

> "In precise terms, we ask two questions:
> - Was there enough freedom available in the conduct of the experiment that a small significance level could have been obtained merely by exploiting it?
> - Is there any evidence for that exploitation?" (Pg. 151)

Concerning the first question they assert that "The first question is answered affirmatively in Section 6…" We strongly disagree with this claim. The work they present in Section 6 is not a scientific work and may only mislead the reader. The answer to this question can only be discussed within the realm of Rabbinical bibliography, and since our present paper deals with the statistical aspects of the debate, we refer the reader to another paper [3] which deals directly with the first question.

Concerning the second question MBBK write:

> "To answer the second question, in Section 7 we examine a very large number of minor variations on WRR's experiment…"

Here they refer to their "study of variations" which is a central part of MBBK's paper. They claim that this study is meant to indirectly prove whether WRR "tuned" the second list of names and appellations to succeed in their Genesis experiment. Our paper will concentrate in critiquing MBBK's "study of variations".

MBBK describe their basic approach as follows:

> "Our method is to study variations on WRR's experiment. We consider many choices made by WRR when they did their experiment, most of them seemingly arbitrary… and see how often these decisions turned out to be favorable to WRR." (Pg. 158)

MBBK know very well that there were no such choices in WRR's second experiment: The second experiment was constrained by the definitions and parameters of their first experiment and there was no room for choices. Therefore MBBK make the following hypothesis:

> "…the apparent tuning of one experimental parameter may in fact be a side-effect of the active tuning of another parameter or parameters.
> For example, the sets of available appellations performing well for two different proximity measures A and B will not generally be the same. Suppose we adopt measure A and select only appellations optimal for that measure. It is likely that some of the appellations thus chosen will be less

> good for measure B, so if we now hold the appellations fixed and change the measure from A to B we can expect the result to get weaker. A suspicious observer might suggest we tuned the measure by trying both A and B and selecting measure A because it worked best, when in truth we may never have even considered measure B. The point is that a parameter of the experiment might be tuned directly, or may come to be optimized as a side-effect of the tuning of some other parameters." (Pg. 159)

To have a scientific meaning, MBBK's "Study of Variations" must be based both on:
(A) A firm (proven) hypothesis;
(B) An unbiased set of independent variations.

Any failure of **(A)** will make the study worthless.
Any failure of **(B)** will nullify not only the study itself, but cast grave doubts on the integrity and honesty of the testers themselves.

Even assuming for argument's sake that MBBK's work has scientific significance, the absence of an objective closed set of variations means that the results of the study have two possible interpretations:

(1) They prove that "tuning" was involved in assembling WRR's data.
(2) They prove that "tuning" was involved in assembling MBBK's variations.

In this paper we will demonstrate that the results of MBBK's "study of variations" indicate not **(1)** but **(2)**.

- Chap. I will list the serious flaws in MBBK's work, both logical and statistical. It should be emphasized that even one such deficiency negates the value of their entire work.
- Chap. II brings examples of serious mathematical-statistical mistakes and deceptions.
- Chap. III shows how MBBK reveal only **part** of the measurements they conducted, and that the way they chose to present those results seriously skews the **true** picture which would be drawn from their own variations. We will also explain the fallacy of MBBK's *a posteriori* excuses for their partial presentation of their results.
- Chap. IV submits their thesis to control experiments. For example, we will examine how their thesis performs on an admittedly "cooked" list — the list they themselves "cooked" to succeed in "War and Peace". This experiment is based on the prediction they themselves allege [4]: That the results of their list for "War and Peace" should worsen and/or improve to the same extent as WRR's list.

> But their prediction **has failed.** The experimental results destroy their thesis: Applying the variations to their list in "War and Peace" worsens the results only in less than half of the variations!

> In this chapter we will bring evidence indicating that MBBK's results of the "study of variations" are due to "tuning" of its variations.

- Chap. V will unfold the "evolution" of the "study of variations". This evolution went through at least four stages. The researchers changed the set of variations time after time, and made ***a posteriori*** changes in the presentation of the results, and each new presentation (even using the same variations) was advantageous to MBBK's goal. In this chapter we will bring further experimental evidence of MBBK's "tuning", and show that there is no connection between their presented results and any "optimization" of data in WRR's work.

CHAPTER I

LOGICAL AND STATISTICAL FLAWS IN MBBK'S THESIS

Let us judge the thesis MBBK invented for their "study of variations", both logically and statistically. Our main criticism centers around the following points.

1. The "study of variations" supposedly helps decide between two possibilities:

A. WRR — Genesis contains hidden ELS codes.

B. MBBK — There is no evidence for hidden codes in Genesis. WRR "tuned" the word list (names and appellations) of the experiment.

If there is indeed a code phenomenon, it must have certain characteristics, and very probably only experiments based on these characteristics will succeed. For example, the construction of the function *c(w, w')* which measures proximities of word pairs, was done in an attempt to incorporate those characteristics observed in earlier examples. The chosen function reflects, for example, the fact that in these earlier examples both parameters *l* and *f* are small (see appendix for chap. I). No doubt this function is not unique, but it is reasonable that other successful functions will exhibit the same characteristics as those detected through the earlier examples.

But if, as MBBK claim, there is no phenomenon, WRR's "tuned" success is geared for only this specific function. So according to them too, the "tuned" list will only succeed with this function or with a similar one.

Therefore, even if for many of the variations the "study of variations" reveals that the results become weaker under the variations, it cannot be determined if this is because of A or B. In both cases the weaker results would be expected, either because the code phenomenon only fits certain characteristics, or because the list was "tuned" according to those characteristics. **This is a fundamental flaw** of the "study of variations".

2. MBBK try to amend this basic flaw through… another fundamental mistake. Let's examine their article:

> **"Regression to the mean?**
> "In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test - and the top group will on average fall back. This is the regression effect." (Freedman, Pisani and

> Purves, 1978). Variations on WRR's experiments, which constitute retest situations, are a case in point. Does this, then, mean that they should show weaker results? If one adopts WRR's null hypothesis, the answer is "yes". In that case, the very low permutation rank they observed is an extreme point in the true (uniform) distribution, and so variations should raise it more often than not. However, under WRR's (implicit) alternative hypothesis, the low permutation rank is not an outlier but a true reflection of some genuine phenomenon. In that case, there is no a priori reason to expect the variations to raise the permutation rank more often than it lowers it." (Pg. 159)

In the emphasized excerpt they claim, that assuming there is a phenomenon,

> "there is no a priori reason to expect the variations to raise the permutation rank more often than it lowers it".

But this is fundamentally wrong as we explained in 1. because a phenomenon would have certain characteristics, and it is highly probable that only an experiment designed according to those characteristics would succeed.

They continue there with the same basic mistake:

> "This is especially obvious if the variation holds fixed those aspects of the experiment which are alleged to contain the phenomenon (the text of Genesis, the concept underlying the list of word pairs and the informal notion of ELS proximity)."

They clearly intimate here that the need for the variation not to deviate from the limits of the phenomenon is **not essential.** It only comes to improve their claim. Had they understood that this condition is essential, rather than merely helpful, two errors would have been avoided:

(A) They give no precise definition of what is included in the phenomenon and what not. By using unclear expressions like—"the concept underlying the list of word pairs and the informal notion of ELS proximity", they easily allow themselves to continue–"Most of our variations will indeed be of that form", because almost anything can now be included in their "definition". This flaw is especially glaring when McKay writes in his report [5] (from where many of the variations were taken):

> "As a qualitative exploration of the set of "reasonable experiments", we examined experiments which are "close by" in the sense that they differ from the original only in some simple way. The classification of these similar experiments as more or less reasonable than the original **is highly subjective**". (Emphasis ours)

(B) One would expect that for such a study, **all** of the variations used should be of this form, and not only **"most"**. Surprisingly, they allowed themselves to include even other types of variations.

We can now understand why MBBK allowed themselves to include variations that deviate from the limits of the phenomenon even according to them. (Actually, unlike them, we think that **most** their variations deviate from the limits of the phenomenon).

At any rate we now have two explicit facts derived from MBBK's own words:

- Some of their variations deviate from the phenomenon as described by WRR, and therefore may be expected **in advance** to damage the results.
- Even the definition of the other variations as "similar to the original experiment" is "highly subjective".

3. MBBK assert in Section 4 of their paper, that WRR's result was improved because of a fluctuation. According to their assertion that a fluctuation did indeed improve the result, it is guaranteed *a priori* that applying the variations will weaken the result more frequently than improve it.

In the next chapter we will furnish an example of how MBBK used this simple fact for their benefit.

4. Their essential underlying assumption of MBBK is that the optimization of the lists (mainly, appellations) should also manifest itself as an optimization of the experiment parameters. But in science and mathematics any assumption must be **proved.** It's amazing that MBBK saw no necessity to prove this assumption. We will see later (in chap. IV) that experiments on MBBK's thesis throw strong doubt on this assumption.

5. MBBK state many times that:

- [they made] "minor variations on WRR's experiment" (Pg. 152).
- "Our approach will be to consider only minimal changes to the experiment." (Pg. 159).
- "However, since almost all the variations we try amount to only small changes in WRR's experiment, we can expect the following property to hold almost always…" (Pg. 159).
- "We believe that in fact we have provided a fairly good coverage of natural minor variations to the experiment and that most qualified persons deeply familiar with the material would choose a similar set. We are happy to test any additional natural minor variation that is brought to our attention." (Pg. 161). (Emphasis ours)

But nowhere is there an *a priori* definition of the terms:
"minor variations", "minimal changes", "small changes", "natural minor variations to the experiment".

With no such criteria anything that follows in necessarily subjective. Indeed we will show later (chap. II) that:
a. Many of their changes were not small and certainly not minimal.
b. In cases where we checked changes smaller than those of MBBK, we got a **completely** different picture.

6. A basic problem with experiments like MBBK's "study of variations" is the **interdependence** of the variations: This interdependence may be between the functions chosen for this purpose, or between the chosen sampling values for a certain parameter. In fact, most of the variations chosen by MBBK have this flaw. As

a direct consequence of these interdependencies, MBBK admit that their results are unquantifiable".
But even though they **cannot quantify their results** they still use them to create a **psychological** impact. See paragraph 10.

Since the name of the game becomes "psychology", MBBK's **presentation of the data** plays a central role**.** Under these circumstances, any misleading presentation of the data has a great impact on the reader. We will give explicit examples of this in chapters II and III.

7. MBBK's list of variations is not closed.

(A) We do not know how many variations were actually attempted, and how many were thrown into the "wastebasket" (MBBK admitted of checking hundreds of variations [6]). MBBK, however, expect us to believe that:

> "Nothing we have chosen to omit tells a story contrary to the story here." (Pg. 152)

But we have already proven [7] [8], and we will prove again in this paper, that it is impossible to rely on this claim. One of the MBBK authors, Prof. Bar-Hillel, publicly admitted (see chap. V, 1(C)(1)) that some variations were indeed thrown into the "waste basket" after using them in arguments against WRR—arguments which were subsequently refuted.

(B) A list which is not closed is insecure not only against hiding unwanted variations, but also against **adding** variations. We mean *a posteriori* examination of variations based on prior knowledge of what will succeed and what will not. In chap. II we shall see many such examples among MBBK's variations. In chapters II and V we will bring statistical evidence that the choice of additions was "tuned" so as to lead to their desired conclusion.

Here is the place to emphasize, that already after the first stage of "evolution of variations", the mathematician Prof. Robert Aumann wrote to Prof. Maya Bar-Hillel as follows [9]:

> "First of all, whatever you do, you've got to say BEFOREHAND "I'm going to do this and that and that."
> You've got to do that BEFORE you actually compute anything. And, you've got to give PRECISE criteria for success and failure. YOU can make them up as you wish, but you've got to tell the world BEFOREHAND what they are. And success or failure, you've got to tell us afterward how your tests came out. So we can keep score.
>
> That's what they did. I didn't believe they would, but they did. And if you want to convince ME, you're going to have to do the same.
>
> If at first you don't succeed, you can keep trying. Just tell us BEFOREHAND what you're doing, and what the criteria are, and whether or not this test is going to be definitive, and so on. You can keep it open, or close it, or do what you want.
> Just tell us. Beforehand."

Nevertheless, MBBK did not fulfill these conditions, which are elementary in any scientific study. Instead they turn to the reader and ask for his trust (page 161):

> "Our selection of variations was in all cases as objective as we could manage; we did not select variations according to how they behaved".

In other words, they say: Believe us that we didn't "tune" the variations, and ***therefore*** don't believe that WRR didn't "tune" their list of appellations.
Therefore the entire "study of variations" is based not on scientific procedure but on faith.

8. In the "study of variations" there is confusion between two different scientific concepts: variations and replications.

(A) Replication is in essence examining the behavior of a phenomenon. We make assumptions about the phenomenon and examine them. For example, in our case, one might assume that the phenomenon occurs not only in Genesis but also in Exodus; that it should occur using other date forms etc. Experiments that are replications change the basic data: for example one may introduce new appellations, new dates, a new book, or a new group of ELSs.

MBBK's investigation could be done using certain assumptions. For example they could assume that if the phenomenon exists, it should also exist in Exodus. However, destroying this assumption would not prove that WRR forged. It would just mean that MBBK's assumption was incorrect. "Why" Genesis differs from Exodus - would be the realm of metaphysics, not statistics.

Therefore any test which is based on assumptions about the nature of the phenomenon is an inappropriate tool for examining the claim of optimization.

(B) In particular, MBBK's "variations" presented in Appendix B of their paper cannot be included in their "study of variations".

Their basic mathematical assumption is that the phenomenon is independent of the specific parameter on which variation is done. In such a case, they assert, they expect that sampling various values for the parameter will (on the average) improve the result as often as it will worsen it.
But such an assumption is incorrect concerning the "variations" presented in their Appendix B, which are based on changes of the data. The phenomenon – which is the existence of a code in the Book of Genesis – is surely expected *a priori* to be dependent on the data. The very nature of tracing a code depends strongly on the data: If we use false data or incorrect linguistic formations, or even if we use acceptable formations but not those included in the code – we will get dramatically different results. Therefore, these "variations" are actually replications and shouldn't be included in the "study of variations".

Most of the MBBK's replications are found in appendix B of their paper and we will deal with them in a separate paper [10].

9. MBBK's choice of four specific statistics with which to present the results of the variations (the section entitled "What measures shall we compare" pg. 160) is an ***a posteriori*** choice of **a part** of the measurements. They chose one quartet which is

one combination out of many millions of possible combinations (see the appendix to chap. I), relying on *a posteriori* reasoning. In principle, this is a fundamental flaw.

Furthermore we will see later that this flaw is not only theoretical, but is an essential component of their study that distorts the results. In chapter II we will show that this *a posteriori* choice allowed significant facts to be concealed. In chap. III we will present the hidden part of the data, and it will be obvious that the true picture shows the opposite of their claims: That even according to MBBK's choice of variations **there was no optimization by WRR**.

10. The "study of variations" lacks quantitative assessment.

MBBK write:

> "For these reasons… we are not going to attempt a quantitative assessment of our evidence. We merely state our case that the evidence is strong and leave it for the reader to judge." (Pg. 159)

But how could a study lacking quantitative assessment be published in a statistics journal!?

In conclusion:

In conducting an investigation of possible optimization by use of the "study of variations", one must:

A. Prove the validity of the working model (see paragraph 4 above).
B. To have a means to distinguish the case of optimization from that of a code or a fluctuation (see 1 and 3 above).
C. Take care that the set of variations stands up to **each** of the following conditions:
   1. It must not deviate from the limits of the phenomenon (see 1, 2 and 5 above).
   2. It must not deviate from the conditions of the original experiment (see 8).
   3. It must be demonstrably closed and a-priori (see 7 and 9 above).
   4. They must be independent (see 6 above).
D. Provide a quantitative assessment of the results (see 10 above).

MBBK's work satisfies **none of these necessary conditions.**

Until now we have shown fundamental logical and statistical flaws in MBBK's "study of variations". It is clear that the study is disqualified even it contains a single one of the fatal flaws enumerated. Indeed, we shall see later in chaps. IV and V, that MBBK's thesis collapses entirely under control experiments.

Before doing this, however, we want to discuss the following: We saw above that MBBK invited the reader to judge the results. But then, a natural question arises: Did MBBK supply the reader with a true picture in order to do so? Did MBBK supply the reader with reliable data, gathered by correct and unbiased sampling and given in an undistorted presentation?

We shall present an answer to this in chapters II and III.

CHAPTER II

MISTAKEN AND MISLEADING SAMPLING IN MBBK'S "STUDY OF VARIATIONS"

This chapter gives examples of erroneous sampling done by MBBK. For the sake of clarity, we have divided the examples into two sections: 1. Misleading sampling. 2. Mistaken sampling. Sec. 1 also includes an example of misrepresentation of data, but this issue is mainly discussed in the next chapter.

1. Misleading sampling:

(A) In table no. 5 (page 169 in their paper), in the middle column, MBBK examine the results of 33 functions which are variations of the function *delta*= the "distance" between two ELSs. All the functions are in the second power like that used by WRR's. The right hand column of that table presents the results of 34 additional variations: One of these is the square root of the original function, and the remaining 33 variations are created by taking the square root of the 33 above mentioned functions. Thus the right hand column has 34 additional functions, all in the first power.

Examining the right hand column of results reveals a striking fact: **All** 68 results for the second list (two results for each variation) are very weakened. And the same occurs to **all** the 68 results of the first list. A combined result of 136 deteriorations versus 0 improvements appears **extremely** improbable even according to MBBK's thesis. Probably, only naive statistical expectations of the outcome (see their discussion of these results on pg. 169) prevented MBBK from noticing that the combined result was "too good" even for their hypothesis.
The solution to this puzzle is simple and astounding:

(1) Changing from the second power to the first power is the dominant cause of the weakening of the results. Therefore the form of the function, from which the square root is taken, has only has a secondary effect on the results:

Comparing the results of the right column to the corresponding results of the middle column, reveals that the former is always much worse. Gans [11] compared the population of results for the variations of the second power with the population of results for the variations of the second power. He did this for P4 (the column third to the left on each row) and he reports:

> "Specifically, the Mann-Whitney Sum of Rank statistic comparing the two populations gives a score of 6.42, indicating that the probability of the two sets of variations coming from the same underlying distribution is 6.8E-11."

Later Gans [12] also compared the remaining three columns and here is his conclusion:

> "Mann – Whitney:
> Column 1: 6.31 sigma, p=1.4E-10.
> Column 2: 6.97 sigma, p=1.6E-12.
> Column 3: 6.42 sigma, p=6.8E-11.

> Column 4: 6.95 sigma, p=1.8E-12.
> All 4 columns together: 12.88 sigma, p=2.9E-38."

(2) Even more. **It is predictable in advance** that **combining** the two variations–– the change to a function of a type that weakens the results, together with a variation which takes its square root (that also weakens the results)—will only damage the results **even more.** MBBK themselves write on pg. 159:

> "However, since almost all the variations we try amount to only small changes in WRR's experiment, we can expect the following property to hold almost always: if changing each of two parameters makes the result worse, changing them both together also makes the result worse."

(3) MBBK knew this in advance, **before** they checked all the 34 functions, and not only theoretically. From tests done by McKay about three years earlier [13] on three functions of the first power, he knew that taking the square root damages the results. Even more, after he published the results for *four* variations of the same kind in CHANCE [4], (together with Bar-Hillel and Bar-Natan), we explicitly replied in that journal [14], that we could have predicted in advance that variations of the first power would destroy the result.

(4) It follows that all the 34 variations to the first power are really the same thing in disguise. Therefore, all 68 results for the second list are really only **2** results. And the same applies to the 68 results for the first list.

It was absolutely unjustified for them to repeat the same basic variation 34 times and present it as 34 variations.

To present the facts as if there are 2 x 68 negative results is a serious deception which calls into question all the variations.

For the non-statistician the following parallel may help:
Galileo reveals Jupiter's four moons for the first time with his telescope. Of course this finding contradicts conventional knowledge, and he is suspected of deceit. His opponents repeat his experiment with many variations of lenses. The result—no moons. Galileo complains: My lens was convex and yours was concave! The use of a convex lens was essential to the experiment and not fortuitous. Your "variation" was incompatible to this experiment. (Similarly, MBBK's use of the first power for the delta function is erroneous). *Galileo's opponents complain that he didn't announce in advance that a concave lens is incompatible.* (Similarly, BBM argued like this in CHANCE). *Galileo's opponents test 33 different concave lenses and still, of course, see nothing. How, they argue, could Galileo have been fortunate enough to see moons when we saw nothing using 34 different lenses?* (Similarly, MBBK argue that they used 34 variations and always got worse results).

The misconception is obvious: Galileo's opponents actually tried only one variation—*the concave lens*. Note: Even had this variation been appropriate, it would still be unjustified to repeat it 33 times and count it as 33 extra variations!

Let us spare MBBK the embarrassment and remove this data. At least the 33 **added** variations of the first power **should not be included in "the study of variations".**

(B) Ignoring the principle of minimality.

(1) Our investigation was based on two main principles. One of them is the principle of minimal skips. This principle is greatly emphasized in all our publications describing the phenomenon. We already wrote in our first pre-print [15]:

> "Our study is based on the following two ideas:
> a. We focus our attention on ELS with minimal skips.
> b. We use two-dimensional arrangement of the text of the Book of Genesis". (Pg. 5)

In addition, we wrote in the *Statistical Science* paper [1]:

> "In Genesis, though, the phenomenon persists when one confines attention to the more "noteworthy" ELS's, that is, those in which the skip $|d|$ is *minimal* over the whole text or over large parts of it." (Pg. 430)

In other words we clearly emphasized the centrality of the principle of minimal skips. This principle has two components:

(i) We claim that the phenomenon is supported by those ELSs which are minimal over large portions of text.
(ii) Our calculations give *more* weight to those ELSs which are "*more*" minimal, i.e. minimal over *larger* portions of text.

(2) But anyone reading MBBK's paper wouldn't even **guess** that the principle of minimality exists. In their introduction (pg. 151) they describe the phenomenon and mention the word convergence in general without mentioning the principle of minimal appearances. However, in appendix A, where they perforce must give a mathematical description of the phenomenon, they do devote several lines (pg. 168) to formally explain the concepts "domain of minimality" and "domain of simultaneous minimality", with no word of explanation of where these previously unmentioned concepts sprang from. (It goes without saying that they never mention the centrality of this principle). These concepts are never mentioned again in the article itself. Even where they point out the central aspects of the experiment, they make no mention of the principle of minimality:

> "This is especially obvious if the variation holds fixed those aspects of the experiment which are alleged to contain the phenomenon (the text of Genesis, the concept underlying the list of word pairs and the informal notion of ELS proximity)." (Pg. 159)

These words clearly ignore the principle of minimality. This becomes even more glaring in light of the fact that McKay himself brings the following quote from our *Statistical Science* paper in his report [5] connected to his variations:

> "We stress that our definition of distance is not unique. Although there are certain general principles (like minimizing the skip d) some of the details can be carried out in other ways. We feel that varying these details is unlikely to affect the results substantially". (Pg. 431, emphasis ours)

One could think that perhaps this omission stemmed from carelessness, and the principle of minimality was actually taken into account. However, we will see in the

next paragraph that this omission allowed the deliberate formulation of variations that **contradict** the minimal principle.

(3) In MBBK's Appendix C, two tables (out of six) of variations, tables 7 and 8, deal with variations closely connected to the minimal principle.

Anyone reading only their paper, ignorant of the vital importance of the principle of minimality, cannot suspect that the "innocent" variation (no. 2 in table 7) is not only neither "natural" nor "minor", but totally ignores a central characteristic of the phenomenon! True, this variation at least uses the minimal ELS's [In accordance to component (i) of the principle of minimality (see (1) above)], but it gives these ELS's no special weight [in contradiction to component (ii) of this principle].

But in the following experiments they did something even worse. They simply "cut off" and discarded the minimal ELS's themselves. They did this with 4 variations at the end of table 8, where they "cut off" from the data precisely the shortest skips, which are part of the minimal ELS's, and threw them into the "waste basket". The same thing happened in the following "simple" experiment (pg. 171):

> "…a simple experiment which to some extent is independent of the original experiment. We did the same computation restricted to those ELS pairs which lie within the cut-off at parameter 20 but no within the cut-off at parameter 10." (Emphasis ours)

But in the mentioned domain between 10 and 20, there remain very few ELS's with "minimal skips in large sections of the book". Therefore, according to our hypothesis, **we ourselves** would have expected failure. MBBK simply "cut off" and discarded almost all the minimal ELS's in large sections of the text. It's as if after Galileo discovered Jupiter's moons, his opponents turned the telescope 180 degrees and announced that they saw nothing! The reader of MBBK's paper would never guess that the above results actually **prove** the WRR's theory by demonstrating that the phenomenon indeed relies on "minimal skips in large sections".
These were just examples. Concerning table 8 see more in the next paragraph.

(C) Let's return to table 8 of their paper, where, in the last four variations, MBBK cut off and discard any ELSs with skips of less than 3, 4, 5, or 10. In the last paragraph we pointed out that these variations are expected **in advance** to be destructive:
(1) Because we know that the code phenomenon relies on minimal ELS's.
(2) And because the discarded ELS's with such short skips, are supposed to contain the main minimal ELS's for several expressions.

But this is not all. MBBK discovered and emphasized that

> "One appellation (out of 102) is so influential that it contributes a factor of 10 to the result by itself." (pg. 155)

The successful appellation referred to is the "Ha'raavi" and it is successful in the second sample. Its success is due to its ELS's with skip 2. McKay knows this well because already in his first report [13] **before** such variations were suggested, he searched for (other) pretexts to get rid of the successful convergences of these "Ha'raavi" ELS's.

It is made clear that he finally found a way to do this through these 4 variations, where he could be **certain in advance** that the results of the second sample would be weakened!

Remark: According to our research hypothesis, we expected to find successful convergences of the minimal ELS's of "Ha'raavi" with its dates. Its main minimal ELS's would be expected in advance to be a skip of 2 (because of the length of the word and the frequencies of its component letters).

(D) We learned from MBBK's paper that their model for "the study of variations" is actually of "losers only".

In their table 5 there are 33 variations which are a function of the second power (as mentioned in (A)). But while using them, MBBK ran into a problem: The P4 statistic (one of the four statistics they chose to present) actually **improved** 19 times out of 33. What did they do? First MBBK joined the 34 variations of the first power to the calculation to lower the percentage of improvements and make it 19 out 67. We showed earlier, in paragraph (A), that this is a deception.

Amusingly, MBBK were still dissatisfied. According to their paper (pg. 169) they had naive statistical expectations of achieving 0 improvements out of 67. Therefore MBBK introduced ***a posteriori*** **pretexts** to nullify even the 19 out of 67 improvements, pretexts published for the first time in their paper, along with the results themselves.

Their main pretext is that the list of appellations of WRR was "cooked" not only for the optimization of P4 but for some other optimization as well, and therefore nothing can be deduced from the improvements of P4. But this story of an additional optimization is not only just another piece of nonsense (see at the end of chap. III) -- but also a mathematical error and a misleading argument.

(1) A mathematical error: Because an additional optimization should not prevent the original P4 from appearing optimized.

(2) A misleading argument: Their statistical model seems something like this:

- If the variations show an optimum for the original results—it proves that WRR optimized the appellations.
- If the variations show no optimum for the original results—it means nothing, because the significance of the results can always be negated by *a posteriori* pretexts.

The police would close any casino working according to such "models".

(E) In chap. I paragraph 9 we pointed out the following basic flaw: Their choice of four specific statistics to present the variation's results (The passage beginning "What measures should we compare" on pg. 160 of their paper) is an *a posteriori* choice of part of the measurements. They chose one quartet from a huge number of other combinations, based on *a posteriori* pretexts. This is a fundamental flaw.

We will devote a complete chapter, chapter 3, to show the complete results of the various statistics, and there we will see that MBBK's choice indeed distorts the picture arising from their own experiments. Here one example will suffice.

Their distorted presentation of the variations reaches the heights of absurdity in the case of "Cut-off defining P1" variations (table 10). Because of the quartet of statistics they chose they need to show the results of P2 (and its analog P4) in an experiment intended to examine the influence of the variations on P1! [These variations were intended to influence P1 (and its analog P3) and indirectly r1 and r3]. They published the following results, giving the data for the 4 chosen statistics:

	L1		L2	
Cut-off defining P1	P2	Min(r1-r4)	P4	Min(r1-r4)
0.05	1	1.0	1	1.0
0.1	1	1.0	1	1.0
0.15	1	1.0	1	1.0
0.2 (WRR)	1	1	1	1
0.25	1	0.8	1	1.0
0.33	1	1.0	1	1.0
0.4	1	1.0	1	1.0
0.5	1	0.4	1	1.0

Table 1

[To make things clearer we added the notation: L1= First list, L2=Second list.]

But if we present the results for the relevant statistics, the picture is as follows:

For L1:

Cut-off defining P1	P1	r_1	P3	r_3
0.05	475487	18.76	134	4.02
0.1	386357	84.42	1205	37.3
0.15	2639	26.13	74	6.43
0.2 (WRR)	1	1	1	1
0.25	0.0024	0.069	0.019	0.13
0.33	0.0008	0.098	2.47	6.12
0.4	0.001	0.19	0.63	4.03
0.5	0.00013	0.036	0.018	0.41

Table 2

For L2:

Cut-off defining P1	P1	r_1	P3	r_3
0.05	105048	18.5	5157	8.04
0.1	133	1.89	6.57	0.26
0.15	145	4.0	14.4	1.26
0.2 (WRR)	1	1	1	1
0.25	0.00032	0.014	0.000015	0.0019
0.33	0.00034	0.05	0.0001	0.018
0.4	0.0083	0.21	0.0048	0.14
0.5	0.055	0.9	0.05	1.0

Table 3

It is important to note that P1 and P2 served as the only statistics to estimate the success of L1 and L2. Therefore, if there was an optimization it was done in relation to P1, or in relation to P2, or -- what is more likely, in relation to Min(P1-P2). Here are the statistics according to the statistic Min(P1-P2):

Cut-off defining P1	Min(P1-P2)	
	L1	L2
0.05	1.32	1.0
0.1	1.32	1.0
0.15	1.32	1.0
0.2 (WRR)	1	1
0.25	0.0024	0.007
0.33	0.0008	0.0074
0.4	0.001	0.18
0.5	0.00013	1.0

Table 4

This table contains an important piece of information. MBBK attempt to cast doubt on WRR's integrity trying to bring indirect 'evidence' that WRR used various manipulations to improve their results. But the variations in this table offer **clear direct proof that WRR worked with complete integrity.** There is nothing easier and simpler than an *a posteriori* choice of the "Cut-off defining P1". A glance in Table 4 reveals that WRR could have improved their results (which were measured by P1 and P2 values) a **thousand fold** by choosing a suitable Cut-off.

But reading MBBK's Table 10 nobody can deduce all this important information. Even encountering what MBBK wrote in their text:

> "Values greater than 0.2 have a dramatic effect on P_1, reducing it by a large factor (especially for the first list). However, the result of the permutation test on P_1 does not improve so much, and for the second list it is never better than that for P_4," (Pg. 171)

does not help too much. How can the reader learn from these words all the data included in our tables 2-4, and how he is supposed to deduce that here we have a **clear direct proof that WRR worked with complete integrity?**

(F) Actually we need not suspect that MBBK always chose the *presentation* worst for WRR. Practically speaking their procedure could have been done inversely: Because according to their thesis they have to manifest optimization for the r-statistics (=the ranks for the permutation test), they could search for (or create, see paragraphs (A), (B), (C) above) specifically those variations that especially worsen these statistics. Since the set of variations is not only "not closed", but wide open, it is possible to search for (or create) such variations quite easily.

For example, in the variations of tables 5-10 one can discern a set A of results for variations of the following type: Variations that are sampling values for parameters or thresholds that were possible in the original experiment. Altogether, MBBK deals with 7 such parameters or thresholds. Because there are several sampling values for each parameter or threshold, there are altogether 44 sampling values divided into 7 groups. MBBK calls each sampling value a "variation".

The calculation of the variations was made in two stages (at least). McKay's report [5] included the subset A1 of the results from A, which contained the results of 22 variations (For L1 the results were for only 20 of these 22 variations). The rest of the measurements took place at a later stage. Let us call the subset of the later results A2=A-A1. At each stage, the choice of sampling values was arbitrary.

Arbitrary choice of sampling values for specific parameters **allows** deception in two ways:

(1) **Deliberate duplication:** In paragraph (A) above, we showed how MBBK multiplied the number of variations in Table 5, while keeping the percentage of weakened results from dropping. They did this by inserting 33 variations of the first power (for details see there). In this case they acted with the same intent: They multiplied the results for the sampling values they had at the first stage (A1), by adding extra results (A2), while keeping the percentage of weakened results from dropping in the second stage.

The technique for this was simple: They sampled again and again from the **same** 7 groups. In other words, after sampling a certain group in the first stage, and finding that most of the results indicated a weakening and only a minority an improvement, they would sample again from that same group even though the results were at that point (more or less) known in advance.

However if we compare the results of the sampling of the two stages, we see that in fact the results of the second stage became even worse, providing MBBK with a desired "covering" to make statements like the following [4]:

> "Wonder of wonders, however, it turns out that almost always (though not quite always) the allegedly blind choices paid off: Just about anything that could have been done differently from how it was actually done would have been detrimental to the list's ranking in the race".

We compare the total of the results (to be detailed in section (2) below, for the 4 statistics chosen by MBBK) in the first stage (A1) to that of the second stage (A2):

	A1	A2
better	12	11
equal	18	14
worse	44	63
total	74	88

Table 5

"better" - Cases where the variation improved result.
"equal" - Cases where the variation made no difference to result.
"worse" - Cases where the variations weakened the result.

There is a clear and remarkable increase in cases for "worse" in the second stage (A2): 43%!

What kind of miracle caused the results of A2 to change so much in comparison to the results of A1?

Now, let us check whether there are some traces to MBBK expectations (or intentions) that many of the variations "would have been detrimental to the list's ranking in the race", that is, to the r-statistics. We define *e*(Ai): the "destructive efficiency" of Ai, to be the average number of "worse" per sampling in Ai.

Sample	L1		L2	
Statistic	P2	Min(r1-r4)	P4	Min(r1-r4)
e(A1)	0.733	0.400	0.706	0.591
e(A2)	0.682	0.667	0.750	0.773

Table 5a

Note the dramatic rise in the "destructive efficiency" of A2 especially in the r-statistics that MBBK prefers!

Such changes for the worse could not have come solely from duplication. We must consider whether there was not some other reason for this. And thus we reach the subject of selective sampling.

(2) **Selective sampling:** Not only is there no set *a priori* method of choosing sampling values (which allows these values to be arbitrarily chosen), but often one can even know in **advance,** from the tendency of the results for certain sampling values, what the results for other values will be. And this can be utilized for selective sampling.

We will now examine in detail the effect of duplication and selective sampling on each of the seven groups of sampling values included in the set A.

(a) In Table 6 of their paper, MBBK present 7 sampling values for "value of *i*": 1, 2, 5, 15, 20, 25 and 50. We will copy this here:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Use 1 value of *i*	2e5	340	31	21
or 2	2e4	210	3.4	4.5
or 5	3.7	0.6	0.3	0.2
or 10 (WRR)	1	1	1	1
or 15	3.6	3.3	1.4	1.1
or 20	11.8	5.9	3.1	3.8
or 25	66	15.3	4.8	5.4
or 50	3600	40	93	28

Table 6

However in McKay's report [5] where he first reported this test, only the values 2, 5, 15, and 20 (emphasized in gray) were measured. Values 25 and 50 are only added **now.** From the results for 10, 15, and 20 one could detect a clear tendency towards weakening of the results. One could guess that sampling 25 and 50 would also weaken the results. Indeed this is what happened. Also comparing the results for 2 and 5 would allow one to guess that for the value 1 one could get worse results and indeed this is what happened. MBBK sampled specifically these points.

However they completely avoided any sampling between 5 and 15, a domain where the weakening of results is **not** certain in advance. And all this despite their declared claim (above chap. I paragraph 5) that they were examining only "minor variations", "minimal changes", and "small changes". Let's do this in their stead:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
i=5	3.7	0.6	0.3	0.2
i=6	2.1	0.5	0.5	0.5
i=7	3.4	2.5	0.3	0.3
i=8	2.7	1.7	0.2	0.2
i=9	0.7	0.7	0.4	0.5
or 10 (WRR)	1	1	1	1
i=11	0.8	0.9	0.6	0.7
i=12	1.1	1.2	0.8	0.8
i=13	1.3	1.3	1.2	1.0
i=14	1.8	2.0	1.1	0.9
i=15	3.6	3.3	1.4	1.4

Table 7

MBBK did not sample between 2 and 5, and in this domain also, there was no advance guarantee from measurements of the first stage, that there would be a weakening of results.

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
i=3	2053	91	1.1	1.8
i=4	119	16.4	0.1	0.2

Table 8

The result most important to MBBK is min(r1-r4) for L2, because this is the result of WRR. In chap. V we will discuss this at length. But here we will just note that in the samplings we added to tables 7 and 8, this statistic improved 8 times, and only worsened once out of ten cases. Compare this to MBBK's results where this statistic improves only 4 times out of 135 variations in tables 5-10 of their paper.

(b) In Table 8 of their paper we find 9 sampling values for “Expected ELS count of": 2, 5, 15, 20, 25, 30, 50, 75, and 100. We will reproduce this here:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Expected ELS count of 2	7600	7.0	4e4	310
or 5	53	53	20	19.5
or 10 (WRR)	1	1	1	1
or 15	1.2	2.9	5.9	2.0
or 20	2.7	8.3	59	7.1
or 25	0.8	4.0	91	15.2
or 30	6.8	14.1	140	22
or 50	2.2	4.1	550	79
or 75	3.7	4.5	590	81
or 100	4.0	4.7	560	62

Table 9

But in McKay's report, results were given only for the values 5 and 15 for L1, and for the values 5, 15, 20, and 30 for L2 (He wrote that the measurements for points 20 and 30 for L1 were “not finished”, and that he “will do 50”). We have emphasized these results in gray.

It transpires that after these results showed a clear tendency to worsen the results for values smaller than 5 and greater than 30, MBBK **later** added the results for 2, 50, 75, and 100. Even more—they allowed themselves to add another sampling value, 25, between the two known (weaker) results for 20 and 30. Thus most of the points were added **after** advance knowledge of the expected tendency.

[*Note: For values 2 and 5, worse results could be expected in advance, because the number of competitors with unequal skips is expected a priori to be smaller than the original (even though this number cannot be known precisely in advance). The effect of altering the number of competitors is itself a reason to weaken the results for values 2 and 5 by a factor of 1.69 and 1.36, respectively, because of this cause alone. An explanation for this can be found in paragraph (e) later. However in this case there are further factors which cause additional weakening*].

Here also they didn’t sample between the values 5 and 15, a domain where worsened results were **not** guaranteed in advance. And this despite their declared stance (see above chap. I paragraph 5) that they would examine only "small changes", "minimal changes", and "minor variations". We did examine these values:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Expected ELS count of 5	53	1.6	20	19.5
Or 6	6.3	0.8	3.8	0.9
Or 7	204	8.8	0.4	0.5
Or 8	6.2	2.4	2.0	0.8
Or 9	9.0	4.1	1.6	1.0
Or 10 (WRR)	1	1	1	1
Or 11	1.3	1.3	1.9	1.8
Or 12	4.7	3.6	1.3	0.7
Or 13	2.4	2.5	4.2	0.9
Or 14	3.0	3.0	3.6	0.9
Or 15	1.2	2.9	5.9	2.0

Table 10

It is interesting that here also the statistic min(r1-r4) for L2 improves a few times. Especially if we remind ourselves that the sum total of this statistic’s improvements in the variations of MBBK in tables 5-10 are only 4 out 135.

(c) In paragraph (E) above we brought the results of variations from Table 10 of MBBK, and we showed the great distortion in the presentation of the data. Let's examine that table again:

	L1		L2	
Cut-off defining P1	P2	Min(r1-r4)	P4	Min(r1-r4)
0.05	1	1.0	1	1.0
0.1	1	1.0	1	1.0
0.15	1	1.0	1	1.0
0.2 (WRR)	1	1	1	1
0.25	1	0.8	1	1.0
0.33	1	1.0	1	1.0
0.4	1	1.0	1	1.0
0.5	1	0.4	1	1.0

Table 11

We notice that until the value 0.2 MBBK sampled every 0.05, but between 0.2 and 0.5 their sampling is sparser. Is this connected to the fact that the domain that yields improvements is specifically the segment (0.2, 0.5], as one would have expected from the histograms published by WRR? Let's sample each 0.05 of this segment:

	L1		L2	
Cut-off defining P1	P2	Min(r1-r4)	P4	Min(r1-r4)
0.05	1	1.0	1	1.0
0.1	1	1.0	1	1.0
0.15	1	1.0	1	1.0
0.2 (WRR)	1	1	1	1
0.25	1	0.8	1	1.0
0.3	1	0.3	1	1.0
0.35	1	0.3	1	1.0
0.4	1	1.0	1	1.0
0.45	1	1.0	1	1.0
0.5	1	0.4	1	1.0

Table 12

We see that we have gained two more improvements for Min(r1-r4) in L1: This is not a bad yield considering that MBBK "allowed" 13 improvements in this statistic for all the 135 variations. If we present the results correctly, that is according to the statistic Min(P1-P2), we receive an improvement for all the points we added: 0.3, 0.35, and 0.45, both in L1 and in L2.

[*Note: In light of this little experiment, one can understand the changes MBBK made in this data, compared to McKay's report. There the segment (0, 0.2] had four sample points: 0.01, 0.02, 0.05, and 0.1, whereas the segment (0.2, 0.5] which is 1.5 times larger had only three sampling points: 0.25, 0.33, and 0.5. In our opinion MBBK attempted to create a more uniform spread of samples, by removing the points 0.01 and 0.02, and adding 0.15 and 0.4. Thus they created an impression of more balanced sampling without having to pay any price: The change was made to retain the same results, while still retaining sparse sampling in the domain expected to show improvement. At any rate, it is worth noting MBBK's* ***free hand*** *in choosing sampling values*].

(d) Another example of sampling concentrated in a domain where the results would be expected to worsen.

The evaluation of convergences between the ELSs (Equidistant Letter sequences) of the expressions *w* and *w'* is made by comparing them to the convergences between the PLSs ("Perturbed" Letter sequences) of the same expressions. A "contest" is set between the ELSs and a group of various PLSs, to see which of them has the most "successful" convergences.

The contest's result is the function *c(w, w')*, and this is a simple fraction *a/m* where

a= the ranking of the convergence between equal skips, and

m= the number of all the competitors.

If *a/m* is close to 0 it means success—the ELSs were ranked in one of first places. If *a/m* is close to 1 it means failure. The ELSs were ranked in one of last places.

In the original experiment, we didn't include cases for which *m* was less than 10. There was a clear reason for this. Imagine if there were very successful convergences of ELSs in Genesis, where the converging ELSs were not only very close but also "rare", so that the odds of them appearing as ELSs by chance would be very low. Because of this low probability, there would be no "competitors" with unequal skips (PLSs): They would simply not appear! The ELSs alone would compete and the result would be 1/1 (because the number of competitors was reduced to one). This result would indicate complete failure! (Remember, the closer the value of *c* approaches 0 the more success and the closer to 1 the more failure). Even if there were one other competitor (with unequal skips), the distortion of results would still be untenable: 1/2 is a value that contraindicates success. To prevent this distortion, we established a threshold of *m*=10 competitors.

MBBK present in table 10 of their paper, a few values for this threshold (denominator bound).

Variation	L1		L2	
Denominator bound	P2	Min(r1-r4)	P4	Min(r1-r4)
2	2.9	1.0	1.0	1.0
3	2.9	1.2	1.0	1.0
4	1.8	1.2	1.0	1.0
5	1.8	1.2	1.0	1.0
10 (WRR)	1	1	1	1
15	1.0	1.0	1.0	1.0
20	1.0	0.9	1.1	1.1
25	1.0	1.0	1.1	1.1

Table 13

Here they sampled the values 2, 3, 4, 5, 15, 20, and 25. But according to McKay's report originally only the values 2, 5, 15, and 20 were sampled (We have emphasized them in gray background, and included WRR's threshold for comparison).

According to what was said above, it was known **in advance** that the threshold of 2 or 3 could have only detrimental effect on the statistics P1-P4 (and the threshold of 4 on P1 and P3, because for them any result greater than 1/5 is a failure), and r1-r4. It is worth noting that specifically in the range of thresholds expected to fail, the sampling is most concentrated: 2, and 3, and also 4! Here is another example of sampling concentrated in a domain where the results would be expected to worsen.

The sampling concerning the value 25 is also instructive. Anyone who read the list of the *c(w, w')* values for L2 in our pre-print from '88 (and MBBK did read it) knows that of all the 163 pairs in L2, only in **one** case was the denominator less than 20: There was one pair whose result was 4/19, and **no pair** had a denominator of 20,

21, 22, 23, 24, or 25. When MBBK sampled the “denominator bound” of 20 they thus erased the result of 4/19, and thus weakened the results as presented by them. Now MBBK added yet **another sampling**: a threshold of 25, whose affect is exactly like that of threshold of 20 – it erases the result of 4/19!

(e) Another example of sampling concentrated in a domain where results are expected to weaken.

As we said in the previous paragraph, the computation of the original function *c(w, w')* is made by "contest" between ELSs and a group of PLSs. In all there were 125 competitors. Let's say the ELSs excelled in a contest for some pairs of expressions, and the result was *c*=1/125. Question: What would happen if instead of 125 competitors there were only 25? Answer: The result would be 1/25, in other words 5 times **worse**. Similarly if there are 49 or 81 competitors the results would be known **in advance** to be worse (Instead of 1/125 we would get 1/49 or 1/81). This would influence statistics P2 and P4 and r2 and r4. We well remember that MBBK (on pg. 155) explained that the influence of the **smallest** values on these statistics is significant.

WRR’s list of the *c(w, w')* values for L2 included five results of 1/125. Even if we freeze all the other values and change only these five so that instead of 125 competitors we take 25, 49 or 81, the statistic P4 for L2 will be worsened by a factor of 19.8, 15.8 and 2.3 respectively!

How does one establish the number of competitors?

We already explained in our original paper that the competitors with unequal skips (PLSs) are created by applying perturbations to the ELSs through three perturbative variables (x, y, z). Each variable can have one of the 5 values: -2, -1, 0, 1, and 2. Thus there are altogether 5x5x5=125 possibilities of perturbing the equal skips and this is the number of competitors.

One can alter the number of competitors in two ways:

- It is possible to change the range of values $2n+1$ for each perturbative variable. If n=3 we have $2n+1=7$ values for each variable and therefore 7x7x7=343 competitors. For n=4 we have $2n+1=9$ values for each variable and therefore 9x9x9=729 competitors, and for n=5 we have 11x11x11=1331 competitors. And the opposite: If we lessen n and take $n=1$ we get only 3x3x3=27 competitors.
- It is possible to alter the number of perturbative variables. For example, instead of 3 variables (x, y, z) we can take just a pair (x, y). And then, if n=2, we will have 5x5=25 competitors, if n=3 we will have 49 competitors, and if n=4 we will have 81 competitors.

Armed with all this information let's examine Table 9 of MBBK where they sample values for possible number of competitors:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Perturb up to 3 places	0.2	2.4	0.04	1.1
or 4 places	0.2	4.2	0.005	0.6
Perturb last 2 places	5*e*4	4.5	6700	38
up to 3 places	118	2.4	340	18.6
or 4 places	2.5	0.6	135	48

Table 14

The sampling result marked in gray is for n=3 and three perturbative variables, which means 343 competitors. It already appeared in McKay's report. But McKay's report has another sampling value: For n=1 and three variables, which means 27 competitors. From its results one learns that indeed, as expected, it is a variation that greatly damages the results. When MBBK added more sampling values in the next stage, they could only add the cases n=4, 5, 6,… which enlarge n. But they were particularly interested in **reducing** the number of competitors from 125: As we explained before this would weaken the results. Therefore, instead of presenting the case n=1 found in McKay's report, they split it into 3 separate results as follows: They changed from three perturbative variables to two variables, and thus they could present results for 25, or 49, or 81 competitors. These are the last three variations in Table 14.

In contrast to the diligence and creativity with which MBBK added sampling values for which there are *less* than 125 competitors, they tired quickly when it came to sampling values for which there were *more* competitors. They sufficed with a sole and single value: n=4 and three variables (second line of table 14). It's a pity, because had they continued such samplings they would have received considerably improved results:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Perturb up to 3 places	0.2	2.4	0.04	1.1
or 4 places	0.2	4.2	0.005	0.6
or 5 places	0.1	5.0	0.0007	0.3
or 6 places	0.07	4.8	0.0003	0.3

Table 15

Here we added *n=5,6.*

(f) In Table 6 of their article, they sampled values 3, 4, 5, and 10 for "Minimum row length":

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Minimum row length of 3	0.9	1.0	1.3	1.2
or 4	0.9	1.0	1.0	1.1
or 5	0.9	1.0	1.2	1.3
or 10	1.1	0.9	5.4	5.9

Table 16

But in McKay's report only the points 1 and 10 were measured, and we were told that 5 was “not done yet”. Thus the results of 1 were omitted, and the results of points 3, 4, and 5 were added. It must be emphasized that WRR's experiment had no parameter like "Minimum row length", as is clear from the definition of H(*d,d'*) in our paper. After the experiment it turned out that because of a "bug" in the program, no tables with row length of one letter were calculated. McKay examined this in his report - this is the sampling value of 1 - and learned that for this value there was indeed no weakening of the result. He admits there that no such threshold was mentioned in our paper.

Yet even though MBBK were aware that there was certainly no intentional threshold here, they chose to conceal this from the reader, and when they describe our measuring method in appendix A of their paper they present the definition of H(*d,d'*) (pg. 167) as if the definition included the threshold Minimum row length=2. And also here, when they present the sampling values they omit the results for the value of 1.

(g) The last group we discuss includes the five last variations in Table 8 of their paper:

Variation	L1		L2	
	P2	Min(r1-r4)	P4	Min(r1-r4)
Minimum skip of 1	1.5	2.1	0.1	5.0
or 3	0.3	0.7	11.1	5.9
or 4	1.2	1.6	16.3	7.9
or 5	0.5	0.8	16.7	11.3
or 10	13.7	0.6	33	35

Table 17

In his report McKay presented the values for the values 3, 4, and 10 which we have marked in gray. Here two more sampling values were added.
In paragraph (C) we explained that variations which "cut off" the ELSs in short skips can be expected **in advance** to be destructive: This problem is shared by all the four last variations of the table, regardless of *when* they were sampled.

All through paragraph (F), while discussing methods of deliberate duplication and selective sampling, we assumed, for argument's sake, that the sampling in the first stage was usually honest. But why assume this? The vast freedom in sampling itself raises many questions.

The fact that sample points were added **later,** and especially where the results could be foreseen, raises the question: **In how many stages** was the data sampled in the original report of McKay?!

In conclusion: The complete freedom employed in choosing the sampling values, together with clear proofs of the utilization of this freedom, makes this sampling statistically worthless.

(G) In chap. I (paragraph 3) we noted, that if the original result was improved by a fluctuation, then applying the variations would usually cause a weakening of that result.

In Sec. 4 of their paper, MBBK describe at length what they consider a fluctuation that improved WRR's original result. Here we will only remark that in any experiment one can define various fluctuations *a posteriori*. The value of this kind of criticism will be discussed elsewhere.

But concerning the present issue, this fluctuation was marked by McKay [13] **before** executing the "study of variations", and therefore MBBK should have taken it into account, while estimating their study's results.

Did MBBK really not notice this point, and take the WRR's result **as is,** without touching the issue of fluctuations? – We can't say so. In Sec.3 of their paper, they note another fluctuation: this time a fluctuation that caused a weakening of WRR's original result.

Let us explain. Diaconis suggested using 1,000,000 random permutations for the decisive experiment, and thus WRR got the result of 4/1,000,000 for min(r1-r4). MBBK liked this result: At the same time that they published on the internet the final version of their *Statistical Science* paper, two of them, Bar-Natan and McKay, published on the same website their paper on their work in "War and Peace" [16], in which they present the min(r1-r4) result of WRR to be 400/100,000,000 (compared to their 57/100,000,000 result in "War and Peace").

But for the purpose of the "study of variations", MBBK take an opposite stand. They assert that using one million permutations was "a sampling error". They cancel the effect of this fluctuation by choosing as a basis for comparison not the value of 400/100,000,000, but 68/100,000,000 instead.

We repeated their variations using the same 1,000,000 permutations as in the original experiment. Doing this we found that by canceling this fluctuation MBBK changed their study's result as follows:

Instead of having: 11 "better", 24 "equal" and 67 "worse",
They now have: 4 "better", 13 "equal" and 85 "worse".

This evidence makes it clear that results of MBBK's study can be strongly biased because of a fluctuation, even one with moderately low probability.

This evidence also makes clear that MBBK used this fact to improve their study's results.

2. Mistaken sampling:

(A) In Table 6 MBBK changed the functions *miu* in several mistaken ways. The most obvious mistake is the changing of the original definition $miu=(delta)^{-1}$ to the definition $miu= -(delta)^2$. Remember, the functional dependence of *delta* on the distance *r* between two ELSs is: $delta \sim r^2$. Therefore $miu \sim 1/r^2$. According to their change, the dependence becomes $miu \sim -r^2$. It is clear that according this new definition, the dominating factor in the evaluation of convergences between the ELSs will become from the more distant ELSs, and not from the closer ones!

The following parallel demonstrates the absurdity of such a change. The law of $1/r^2$ allows one to examine the local effect of the molecules of some chemical reagent in a test tube on a metal atom in a hemoglobin molecule. Clearly the influence of any atom lessens drastically the further away it is placed. Thus we can

safely investigate the local chemistry and physics in the test tube, without worrying about effects of distant molecules like those on Mars.
Let's make a "slight" change (as MBBK would put it) to this law and change it to $-r^2$. Now the influence of Mars becomes the dominating factor and the influence of the local molecules becomes negligible!
Only extreme bias on MBBK's part could have allowed such a mistake. It is probable that never in history has such an error been published in a scientific journal. Our remarks here apply equally to the changes to *miu=−(delta)* and *miu=−ln(delta)* found in that table.

(B) In our first preprint ('86) we emphasized the importance of <u>two</u> elements in the geometrical convergence between two ELSs: Each ELS must be "concentrated" on the two dimensional table (or cylinder). In other words they must have a "small localization parameter" (small *f*), and they should be close to one another (small *l*). See pages 8-9 and 29-30 there. But MBBK ignored this in a sizable number of the variations of Table 5.

(C) The final result of WRR is the outcome of an accumulation of *c* values for all word pairs include in the sample. Therefore it is **expected in advance** that cutting the number of word pairs in the sample will cause a drop in the significance of the remaining group of pairs. As an extreme example: If we remove half the pairs (randomly), we could expect to damage the result a thousand fold. Therefore variations that lessen the number of pairs are expected to weaken the result..
But MBBK made at least six variations in the tables in appendix C that would be expected to cause a drop in the number of pairs:

- The two first variations in their Table 8 (because of the low skip threshold not all of the pairs will appear as ELSs).
- Variations 5-6 in their Table 9; the drop in the number of competitors caused many pairs to have insufficient (less than 10) competitors, and they were not included.
- Variations 6-7 in Table 10 where the high threshold of competitors caused a smaller number of pairs.

(D) According to MBBK an optimization of the appellations manifests itself as an optimization of the parameters. MBBK did variations of various parameters in order to prove that the results show optimization of the parameters.

Let X be such a parameter. We denote with X_0 the value of X used by WRR, and with Y_0 the result of WRR for X_0. In order to examine whether Y_0 is an optimum of the function Y(X), one should take sampling points X_i in the neighborhood of X_0, and to check if Y_0 has an optimal value as compared to the values Y_i. A necessary and elementary condition for such examination is that the points X_i should be taken <u>on both sides of X_0</u>. For instance, it is quite clear that if the function Y(X) is monotonic in the neighborhood of X_0, sampling only those points which satisfy $X_i < X_0$ (or only those which satisfy $X_i > X_0$) will always worsen the result (or always improve it), although Y_0 is not optimum!

Therefore, MBBK should have demonstrated that the points they chose to sample were indeed taken on "both sides" of the values of the original experiment. But they didn't do that. In most of their variations it isn't even possible. For example, in almost all of the variations not concerning a numerical parameter, but rather a

change of the shape of WRR's function, it can't be decided whether they are on "both sides" of the original function.

(E) We have already emphasized in chap. I (paragraph 6), that one of the main defects of MBBK's collection of variations is the existence of dependencies within certain sets of variations. There are even cases in which the dependency is so strong that the whole set of variations should be considered as a single variation.

For instance, in the case that the function Y(X) (see the previous paragraph) is monotonic on one side of X_0, knowing that the result of one sampling point is worse (or better) than the original, enables us to know the other points' results in advance.

In conclusion:

This chapter has included blatant examples of mistaken and misleading sampling.

(1) Variations affected by the defects mentioned here are very many: over 100 variations out of 135. Many of them suffer several defects. All of this is in addition to the basic flaw (we already mentioned in chap. I) common to all of the variations: The collection of variations is not closed, and this allows "tuning" (as we will show in chap. IV).

(2) The examples of mistaken and misleading sampling we have presented illustrate the 'logic', 'judgement' and 'honesty' used in the choice (or creation) of variations used for the "study of variations". In light of this there is little basis for MBBK's appeal to the reader's trust:

> "Our selection of variations was in all cases as objective as we could manage; we did not select variations according to how they behaved". (Pg. 161)

In the next chapter we will discuss another basic component of MBBK's "study of variations: the misrepresentation of data.

CHAPTER III

MISREPRESENTATION OF DATA IN MBBK'S "STUDY OF VARIATIONS"

We have shown in the previous chapter that the choice (creation) of variations by MBBK was highly defective. Here we will discuss the way they presented their variations' results.

Appendix C of their paper contains an impressive series of tables showing the results of the "study of variations" (tables 5-10 of their paper). We have already noted (in chap. I), that even MBBK admit that they can't quantify their results. In the absence of quantification the only value of the "study of variations" is psychological: persuasion of the reader by means of *"impression",* i.e. by the way the results are presented. In this sense, tables 5-10 which are the "show window" of the "study of variations", have a special importance. It is true there are few additional variations scattered here and there through the text of Appendix C (although, some of them do not deserve the name "variation", since they measure something else). But our aim is to investigate the "show window" they presented to the reader, and to show *how* the desired *"impression"* was achieved.

MBBK used four statistics for the representation of their results: P2 and min(r1-r4) for WRR's first list (L1), and P4 and min(r1-r4) for WRR's second list (L2). [*ri* is the rank of *Pi* in the permutation test].

In Chapter I paragraph 9 we pointed out the following basic flaw: MBBK's choice of four specific statistics to present the results for their variations (section "What measures should we compare" pg. 160 in their paper), is an ***a posteriori*** choice of only part of these results. They chose one quartet out of many millions of possible combinations, relying on *a posteriori* pretexts. We will now show that besides being a basic flaw in principle, it is an absolute distortion.

In the first section of the chapter we will give the complete results of the various statistics and show that MBBK's choice seriously distorts the picture arising from their own experiments.

In the second section we will examine MBBK's *a posteriori* pretexts, and see that besides the basic flaw that they are *a posteriori*, they are also invalid.

1. Presentation of the data:

MBBK ask in their heading: **"What measures should we compare?"** This question might be appropriate *a priori*. But how can investigators ask *a posteriori,* after their tests are done, which results should be revealed. The answer should be clear: Present **all** the results! However, since MBBK did not do this, we ourselves went over all the measurements and we present the results.

In tables 5-10 of their paper, 135 different variations are listed. We found a group of 33 variations of table 5 "illegal" (because they all repeat one single variation: taking the square root of *delta,* as explained in chap. II, in the section on "misleading sampling", end of paragraph (A)), and consequently we erased them. So 102 variations remain. Note that the 7 last variations of MBBK's Table 10 relate, according to their definition, only to P1 (or P3) and not to P2 (or P4). So for P2 (or P4) there are only 95 variations.

(A) The true results:

Let us see what happens if we make **the natural choice** according to their thesis: P1 and P2 were the sole statistics used to evaluate the success of the first rabbis list (L1) and the second rabbis list (L2). Therefore, any optimization would be in relation to P1 or P2, or more likely, in relation to min(P1-P2). Thus the natural choice is to examine the picture in relation to these values. The results are:

	L1			L2		
	P1	P2	Min(P1-P2)	P1	P2	Min(P1-P2)
better	35	13	38	35	38	42
equal	10	3	10	21	6	10
worse	57	79	54	46	51	50
not worse	45	16	48	56	44	52
total	102	95	102	102	95	102

Table 18

"better" - Cases where the variation improved result.
"equal" - Cases where the variation made no difference to result.

“worse” - The cases where the variations weakened the result.
“equal” + “better” = “not worse” - Cases where the variations did not weaken the result.

Conclusion:
In both samples the results for **Min(P1-P2)** show **no sign of optimization.**

Remember that the study of variations was originally meant to examine whether there was direct optimization of the parameters (see chap. V, 1(C)). Therefore, non-presentation of these results by MBBK is extremely puzzling. We claim that all the alternative presentations of variation results chosen by MBBK, with all the attendant rationales, hide this basic fact, as we will show later.

(B) Presentation of the results in the "study of variations":

We will now give the complete results of the various statistics, and show that MBBK's choices indeed conceal the true results given in the previous paragraph.

(1) MBBK allow presentation of statistics P3 and P4, even though they were not the measure for the overall significance of the samples. Let's see the results for L1:

	P1	P2	P3	P4	Min(P1-P4)	Min(P1-P2)
better	35	13	18	17	38	38
equal	10	3	21	7	10	10
worse	57	79	63	71	54	54
not worse	45	16	39	24	48	48
total	102	95	102	95	102	102

Table 19

From these six possible results, MBBK chose the one best for them, P2, because its "better" value is the smallest and its "worse" value is the largest. Now let's see the results for L2:

	P1	P2	P3	P4	Min(P1-P4)	Min(P1-P2)
better	35	38	52	31	42	42
equal	21	6	14	7	10	10
worse	46	51	36	57	50	50
not worse	56	44	66	38	52	52
total	102	95	102	95	102	102

Table 20

From these six possibilities MBBK chose… you guessed it, the result best for them: P4. Its "better" value is the smallest and its "worse" value is the largest!
Discussion of the results:
(a) Please, look at the results for L2, the manifest object of MBBK's "study of variations" (see chap. V): There is **no indication of optimization.**

(b) For L1 there are conflicting trends: On the one hand for P1, Min(P1-P4), and Min(P1-P2), there is no indication of optimization. On the other hand, using MBBK's model, there is such indication for P2, P3, and P4. Note that P3 and P4 were first

defined long after the experiment with L1. So it is very strange that specifically for them there is indication of optimization, while for Min(P1-P2) which was the sole criteria for success—there is no indication of optimization! Further discussion of these contradictions will be held in Chaps. IV and V.

(2) MBBK's "study of variations" gives the most weight to the results of the r-statistics (the ranks in the permutation test). Thus they examined the variations not through P1 and P2, which were the sole statistics estimating the success of the original samples, but through the permutation test which was only conceived **two years after** the (alleged) optimizations were made. We will discuss their *a posteriori* pretexts for such a strange and unnatural decision in the second section of the chapter. Meanwhile, we will give a complete picture of the variations' results in the r-statistics. The rank in the permutation test is denoted as r_i.

For L1:

	r1	r2	r3	r4	Min(r1-r4)	Min(r1-r2)
better	31	8	27	6	13	13
equal	10	10	6	14	14	14
worse	61	77	69	75	75	75
not worse	41	18	33	20	27	27
total	102	95	102	95	102	102

Table 21

For L2:

	r1	r2	r3	r4	Min(r1-r4)	Min(r1-r2)
better	32	6	53	4	4	6
equal	11	7	11	6	13	14
worse	59	82	38	85	85	82
not worse	43	13	64	10	17	20
total	102	95	102	95	102	102

Table 22

For L2, MBBK once again chose the result best for them: Min(r1-r4), and for L1 they preferred Min(r1-r4) over r2 for compatibility with the choice for L2. (They couldn't do this for P4, because P4 was not defined at all for L1).

Discussion:

There is no doubt that the results of the r-statistics show fewer improvements than the P-statistics. But the results raise many questions:

- Note the results for r2 in the two samples: Their similarity is striking! Even though, not only are the samples different and built from different word pairs, but also (according to MBBK) the methods of optimization differed: For L1 there was optimization of the parameters themselves, and also of the data and all details of the experiment. But for L2 all the parameters were already established and the optimization could have concentrated only on the data.

- The similarity between the results for r4 in the two samples is even more surprising. Especially if we remember that for L1, the partial group of appellations used for the measuring of r3 and r4 was not even defined in the original experiment.

- Furthermore, the result trends are mixed, most noticeably in L2: On the one hand r3 gives a clear majority to improvements (supposedly proof of non optimization) while on the other hand r4 achieves a record opposite: almost no improvements (supposedly proof of optimization).
  This arouses suspicion that the results have no connection to the existence or otherwise of optimization! We will deal with this in chap. IV.

In light of all these data it is worth reading MBBK's words:

> "Conclusions.
> As can be seen from the Appendices, the results are remarkably consistent: only a small fraction of variations made WRR's result stronger and then usually by only a small amount." (Pg. 161)

Don't be confused! "The results are remarkably consistent", means **the results that they choose to show us.** Not, for example, the actual results in subsection (A), or the overall results seen in this section of the chapter.

2. The Pretexts:

MBBK supply a sequence of *a posteriori* pretexts to justify their choice of what was presented, especially on pg. 160 of their paper under the heading:

> **"What measures should we compare?"**

But after all their casuistry and pretexts two obvious basic questions remain, and they are ignored by MBBK:

- Why don't they give the reader **all** the results?
- Why, contrary to the demands of statistical research, do they not adhere to the measures they had already chosen at a previous stage, instead of changing them *a posteriori?*

Thus even if valid excuses were given by MBBK, it would be disallowed here, because at most they may justify *a priori* choices of what should be presented. But there is no excuse for *a posteriori* choices.

In truth, we think that MBBK's pretexts themselves are incorrect. We will now discuss them, examine their validity, and try to trace what led to their conception.

(A) The choices of P2 vs. P1, and P4 vs. P3:

Let's see how MBBK explain *a posteriori* their choice of P2 for L1 and P4 for L2. They ask:

> **"What measures should we compare?**
> Another technical problem concerns the comparison of two variations. Should we use the success measures employed by WRR at the time they compiled the data, or those later adopted for publication?"

And they reply:

> "In the case of the first list, the only overall measures of success used by WRR were P2 and their P1-precursor (see Section 3). The relative behavior of P1 on slightly different metrics depends only on a handful of c(w, w') values close to 0.2, and thus only on a handful of appellations. By contrast,

> P2 depends continuously on all of the c(w, w') values, so it should make a more sensitive indicator of tuning. Thus, we will use P2 for the first list."

MBBK give the same reply for disqualifying P3 as a measure for L2:

> "For the second list, P3 is ruled out for the same lack of sensitivity as P1, leaving us to choose between P2 and P4."

(1**)** The reader will certainly be surprised at this *a posteriori* pretext if he remembers that MBBK argued **exactly the opposite** earlier (pg. 155 in their paper):

> **"Sensitivity to a small part of the data.**
> A worrisome aspect of WRR's method is its reliance on multiplication of small numbers. The values of P2 and P4 are highly sensitive to the values of the few smallest distances, and this problem is exacerbated by the positive correlation between c(w, w') values. Due in part to this property, WRR's result relies heavily on only a small part of their data." (Emphasis ours).

Is there any end to the acrobatics of seeking pretexts *a posteriori*?

(2) Had MBBK wished to prefer P1 over P2 they could easily have argued the opposite: the claim that P2 depends on all the values of c(w, w') is good reason to prefer P1 over P2. Why? According to their model, optimization certainly introduced the "successful" pairs to the list, and not the "failures". Therefore P2 is less sensitive: it will also be influenced by changes of the "failures" (due to the variations), that have no connection to the optimization. On the other hand, the changes in P1 are generally caused by the changes of the "successful" ones. Incidentally, later in Paragraph **(B)**, you will find a similar argument raised by MBBK to justify the preference of P4 over P2…

What forced MBBK to choose their argument and not the opposite? And what forced them **to choose** at all?

Especially later, in paragraph **(C)**, MBBK labor "to capture tuning towards the objectives mentioned in the previous paragraph", even though these "objectives" are taken from the realm of imagination. Why were they so keen to abandon the opportunity "to capture tuning towards" P1 which was an authentic measure of the lists' success?

(3**)** It is hard to see why MBBK work so hard to disqualify P1 and P3. Even assuming that they are "less sensitive" to variations, **they would still have to show optimization for L1 and L2!**

However, looking at tables 19-20 above shows **why** MBBK have to disqualify P1 and P3:

	L1		L2	
	P1	P3	P1	P3
better	35	18	35	52
equal	10	21	21	14
worse	57	63	46	36
not worse	45	39	56	66
total	102	102	102	102

Table 23

Obviously—because there is no optimum!

(B) The choice of P4 vs. P2:

MBBK continue to explain:

> "These two measures differ only in whether appellations of the form "Rabbi X" are included (P2) or not (P4). However, experimental parameters not subject to choice cannot be involved in tuning, and because the "Rabbi X" appellations were forced on WRR by their prior use in the first list, we can expect P4 to be a more sensitive indicator of tuning than P2. Thus, we will use P4."

MBBK present an amazing argument: They claim that the part of the sample built from the standard appellations "Rabbi X", was subjected to no direct or indirect optimization.

This argument is astounding not because it is incorrect, but because it is a complete contradiction to the impression given by their whole paper!

Let us explain: MBBK explain and explain how much freedom we had to include or exclude personalities in L2, and how much latitude we had to make selective adjustment of the dates. MBBK's "War and Peace" list is built on **three** optimizations: 1. Regarding the appellations. 2. Regarding the inclusion of personalities in the list. 3. Regarding the amendment/omitting/addition of dates.

They justified their three optimizations by claiming [16] that they did exactly like WRR.

But the claim about optimizations 2 and 3 is very connected to that part of "Rabbi X": There are even cases where this claim is relevant **only** to that part of L2!

What brought them to this contradiction? - Two simple facts:

- Correcting the list of personalities according to MBBK's criterion [17] and correction/ addition/deletion of dates (according to their expert), would bring about an overall **improvement** of the original result for L2, i.e. min(P1-P2) by a factor of 3.4 [even if we do not include Rabbi David Ganz in the list, according to MBBK's dubious argument, there is still an improvement by a factor of 1.8].
  The implication of this fact is that WRR did no optimization of types 2 or 3.
- If the group "Rabbi X" were subjected to the variations, it would indeed reveal that there was no optimization (We will bring the results in chap. IV paragraph 3(B)).

But instead of presenting these results publicly, MBBK chose an alternative: They chose to conceal the truth from the reader, and instead create the impression that WRR did such optimizations. The reason: When MBBK "cooked" the "War and Peace" list, they failed to reach the same level of significance as WRR's list through optimization of appellations alone. To do this they *also* had to use optimizations 2 and 3 in order to improve their result by one order of magnitude.

On the other hand, the "Rabbi X" group interfered with MBBK's reaching their desired results for "the study of variations", so to justify its removal they wrote the excerpt quoted above.

(C) The choice of min(r1-r4):

At the next stage MBBK wish to justify, *a posteriori*, their belated use of min(r1-r4). They continue their paper as follows:

> "In addition to P2 for the first list and P4 for the second, we will show the effect of experiment variations on the least of the permutation ranks of P1-P4. This is not only the sole success measure presented in WRR94, but there are other good reasons. The permutation rank of P4, for example, is a version of P4 which has been "normalized" in a way that makes sense in the case of experimental variations that change the number of distances, or variations that tend to uniformly move distances in the same direction. For this reason, the permutation rank of P4 should often be a more reliable indicator of tuning than P4 itself. The permutation rank also to some extent measures P1-P4 for both the identity permutation and one or more cyclic shifts, so it might tend to capture tuning towards the objectives mentioned in the previous paragraph. (Recall from Section 3 that WRR had been asked to investigate a "randomly chosen" cyclic shift.)"

There are three arguments here:

Argument 1:

> "… the sole success measure presented in WRR94..."

This argument is not only irrelevant, but also bizarre: As we already mentioned above, P1 and P2 served as the sole statistics to evaluate the success of the original samples. Therefore, any optimization must have been in relation to P1 or P2, or—even more likely—in relation to Min(P1-P2). Therefore, *a priori*, **the natural choice** would be to examine the situation regarding these values.

Argument 2:

> "The permutation rank of P4, for example, is a version of P4 which has been "normalized" in a way that makes sense in the case of experimental variations that change the number of distances, or variations that tend to uniformly move distances in the same direction. For this reason, the permutation rank of P4 should often be a more reliable indicator of tuning than P4 itself."

This is a typical *a posteriori* argument of MBBK. It actually includes three arguments, one of general nature and two specific:

(1) The general argument is that it is more correct to examine the influence of the variations on r4 values, i.e, the values received through the permutation test, because r4 is "more reliable" than P4. This is a mistake. This argument is perhaps relevant to an experiment checking the quality of the result, but not in the "study of variations" which examines the stability of the results. Obviously, if I suspect that P4 was optimized, it is worthwhile to examine the stability of P4, because over there will be found the maximum sensitivity to change (and at this stage MBBK argued that they were seeking the maximum sensitivity…). On the other hand, it is not at all clear how an optimization of P4 would be manifested in its (complicated) transform, r4. It is feasible that the permutations themselves could destroy part of the (claimed) optimizations, as we will illustrate later (paragraph (3)).

In chap. V we will bring experimental evidence that the variations often damage the results because of their influence on some features of r4, the complicated transform of P4, and not because of any connection to the original experiment.

(2) In the above quote they claimed that r4 is superior to P4

> "in the case of experimental variations that change the number of distances".

[*Note: This claim was first raised by Prof. Gil Kalai In November '97. The sequence of events was as follows: Prof. Bar-Hillel used the results of permutation tests to test the 13 choices she presented in Jan. '97. Prof Aumann [9] criticized the fact that she had calculated the "advantage" or "disadvantage" of the choices, according to the permutations test which was only suggested about two years after the choices were made, and not with the original statistics, P1 and P2 which were used in the original experiment, and which WRR had (supposedly) used to execute their deception. Ten months later Prof Kalai tried to extricate Bar-Hillel from her predicament, and explained how, despite the criticism, she was right to use the later permutation test as a measure for her study of choices. According to our records [18], Bar-Hillel had difficulty understanding Kalai's retroactive explanation, and in fact could not accept it, because she had already announced that she had a different position altogether.*]

(a) In subsection (b) we will explain what this argument entails. Here we will examine if this argument is at all relevant to the variations discussed in this paper, which are those checked by MBBK and specified in tables 5-10 in appendix C of their paper.

From their argument, it is clearly only relevant to variations where the number of pairs (i.e, the number of "distances") changes in the sample. The variations where the number of pairs changes can be divided into two:

(i) Variations where the number of pairs lessened.
(ii) Variations where the number of pairs grew.

(i) Concerning the variations where the number of pairs lessened: As we said in chap. II (section 2, paragraph (C)), there are altogether 6 such variations. As we explained there, it is a mistake to use variations where the

number of pairs is lessened, because they are expected in advance to weaken the results for P4 and also r4. Therefore MBBK should not have used such variations.

(ii) Concerning the variations where the number of pairs increased: MBBK checked very few such variations:

- For L1: The number of pairs increased only in 5 of 135 variations in tables 5-10. It turns out that in these few cases, the addition of these pairs already causes weakening in P4, therefore, even according to Kalai [19] it is pointless to check in r4, because his argument deals with cases where P4 improves and r4 weakens!

- For L2: There are no such variations.

Therefore this argument is totally irrelevant for the variations in tables 5-10 of their paper!

In conclusion: There is no justification to prefer r4 over P4 because of this argument.

(b) We will now explain the argument itself (which is Kalai's argument [19]). The argument is based on a certain feature of some expressions in ELSs, which we called "charisma" [20]. Such expressions have an "advantage" in making successful convergences with others.

Kalai's argument [19] is that (the claimed) optimization has two components:

I. Choosing "charismatic" appellations (Which tend to "succeed" with any other word).

II. Specific optimization concerning convergences with the correct dates.

Upon these assumptions he deals with the following case:

We add a new group of dates to the original group of dates in the sample, so that the "charismatic" appellations of the sample now interact with the new dates and take part in the added "appellation-date" pairs. This results in an improvement in P4. Can one conclude from this that there was no optimization of appellations in the original sample?

According to Kalai, the improvement in P4 is because of I. Therefore, despite the improvement, we cannot conclude that there was no optimization.

His conclusion is, that in cases where pairs are added, we must use the permutation test that nullifies the charismatic effect (because "charismatic" words will succeed even with incorrect dates), or at least radically weakens it. Thus, according to him r4 is better in these cases.

This is the explanation of Kalai's argument.

But, as we explained in (a), this argument is irrelevant for the variations discussed in this paper. On the other hand, it is relevant for replications (in appendix B in their article) where the number of pairs (usually) changes. Therefore, we will discuss this argument in our paper [10] dealing with the replications (in appendix B in their article). There we will bring counter examples to this argument.

(3) MBBK continue and claim that r4 is also better than P4 in the case of variations

"that tend to uniformly move distances in the same direction."

With these words they also refer to the feature we called "charisma". Remember that words with this feature have an "advantage" in convergences. Now, while argument (2) related to variations caused by adding pairs to the original sample, argument **(3)** relates to variations where the number of pairs remains **unchanged**.

This argument says that r4 is better than P4 for variations that increase the charismatic affect.

(a) It turns out that specifically according to MBBK's model, this argument is incorrect, **and the opposite is true.**

Let's examine two arguments of MBBK:

Argument (i) (of Kalai) The optimization (they claim) consists of two components:
I. Choosing "charismatic" appellations (Which tend to "succeed" with any other word).
II. Specific optimization concerning convergences with the correct dates.

Argument (ii): Optimization of appellations is equivalent to optimization of parameters.

Argument (i) says that there was indeed optimization through choice of charismatic words—the optimization of Type I.

From argument (ii) it follows that a variation causing more “charisma” (and thus an improvement in P4) is equal to the choice of more charismatic appellations.

Conclusion: Specifically P4 is *more* sensitive indicator of “type I” optimization. On the other hand, if the permutation test nullifies the charismatic effect, then r4 **can definitely not be an indicator of “Type I” optimization! (**This conclusion holds for all P-statistics versus r-statistics).

(b) Furthermore: Even if, for some reason, we wish to waive the manifestation of “type I” optimization, the method suggested by MBBK is incorrect. The use of the complicated transform r4 for this greatly distorts the results (in chap. V we have experimental evidence for this). On the other hand, there is a simple and correct way to neutralize the charismatic effect of the appellations. In chap. V paragraph 2(B) we will discuss the results of such an experiment. It turns out that the results are totally different to those obtained by transforming to r4.

(c) This new (and mistaken) argument was **first** raised in MBBK’s article in *Stat. Sci.* Perhaps it is a strange coincidence that only at this stage of the "evolution of variations" does this argument bring them some benefit. They use it on pgs. 169-170 to justify the erasure of **19 improvements** of P4 (out of 33 variations) in table 5. They write:

> *“Furthermore, in all 19 cases where P4 dropped, the permutation rank of P4 increased. This indicates that the observed drop in P4 values is due to an overall tendency for c(w, w’) values to decrease when these variations are applied.”*

Note that this phrase can be easily *inverted* regarding these 19 variations:

> Furthermore, in all 19 cases where the permutation rank of P4 increased, P4 dropped. This indicates that the observed increase in the values of the permutation rank of P4 is due to an overall tendency for permutation ranks to increase when these variations are applied.

And MBBK's conclusion:

> *"in other words, it is an example of the inadequacy of P4 as an indirect indicator of tuning, as discussed in Section 7,"*

which we showed above to be **incorrect**, could be turned around to form a more logical conclusion, which stems from the *inverted phrase*:

> In other words, it is an example of variations being chosen according to their destructive effect on r4, as discussed in chap. V (of this paper).

Argument 3:

MBBK's third argument (pg. 160) for preferring r4 over P4 is as follows:

> "The permutation rank also to some extent measures P1_4 for both the identity permutation and one or more cyclic shifts, so it might tend to capture tuning towards the objectives mentioned in the previous paragraph. (Recall from Section 3 that WRR had been asked to investigate a "randomly chosen" cyclic shift.)"

Our reply:

(1) This argument too, is a basic mistake: a further optimization would not necessarily prevent the original P4 from appearing as an optimum! On the other hand it is not at all clear how the optimization on P4 would be expressed in its (complicated) transform, r4.

(2) Even according to them, this *a posteriori* pretext applies only to the second list. For the first list no cyclical permutation was performed! Therefore, if the measurement of the two optimizations requires **different** tools than the measuring of one optimization, there is no reason to use r4, which according to them is suitable to measure two optimizations, for the first list.

(3) Actually, MBBK could argue the **opposite**: To prefer P4 over r4. They write in Sec. 8 of their paper that we made an additional optimization in the second list to get a P2-value very similar to the P2-value of the first list. If so one should especially examine the variations with P4 and not with r4 which is unsuitable "to capture tuning towards this objective". The only disadvantage to this opposite claim is that it leads to results that show no optimization…

(4) **Now for the facts:** Of what additional optimization does MBBK now speak? According to them we made sure that for a certain cyclic permutation there should be "a large value of P2 or P4" (pg. 160).

It turns out that for P4 of that cyclic permutation, there is no optimal value even among the 31 possible cyclic permutations: It is the third largest. The probability for this is about 1/10. Actually, if we experiment we will find that its rank among other permutations of its kind (permutations that have no intersection with the identity permutation), is 816 out 1,000. If so, we have **an *a posteriori* observation of an event whose probability is 0.184—and this is *one* observation out of many possible *a posteriori* observations—that is all!**

What's astounding is, that based on this illusion, MBBK build an entire argument to justify their *a posteriori* preference of r4!

IN CONCLUSION:

- All the justifications MBBK brought for the strange decision not to use the natural choice, which would be the original statistics for which (they say) the optimization was done—were to enable a distorted presentation of the results of the variations.
- All of these justifications have been refuted.
- It should be emphasized that even according to them, they had no right to change the presentation used in the first overall report of McKay [5]; in other words they should have presented the results for P1, P2, r1, r2.
- It should be emphasized that even according to their model, the original measures of success should have shown optimization. But, as we show in this chapter, there is no indication of optimization in the original measures or in several statistics that MBBK preferred not to present.

In the following chapters we will bring more experimental evidence that MBBK's variations were tuned to achieve exceptional results for r2 and especially for r4, results intended to "incriminate" WRR in the optimization of their data.

CHAPTER IV

EXPERIMENTS THAT REFUTE THE "STUDY OF VARIATIONS"

Now we will describe a number of experiments and calculations we performed to test MBBK's thesis. We will discuss the conclusions of each experiment, and finally decide whether the overall emergent picture (including the results presented in chap. III) supports MBBK's thesis or contradicts it.

1. MBBK's prediction:

In an earlier article in the *Chance* journal [4], Bar-Hillel, Bar-Natan and McKay first presented the "Study of Variations" thesis in it's present guise, and also put forward a prediction:

> "Lest there be a misunderstanding, we hasten to repeat that the fact that a particular choice made by Witztum and Rips turned out to be better than its alternative by no means implies that both were checked and the superior one was chosen. The method whereby War and Peace list is cooked did not involve any of these choices, because they were imposed already. All choices were limited to which names and appellations to include and how to spell them. Nonetheless, our

> list would have fared similarly to theirs under the same checks. If a list of names is cooked to optimize some statistic given some choices, the choices look as if they were cooked to optimize the statistic given the list of names." (Pg. 19, emphasis ours)

As we already noted (chap. I sec. 4), MBBK's claim that the optimization of the lists should also manifest itself as an optimization of the experiment parameters, is no more than an *assumption*. But in science and mathematics such assumption must be **proven.** It's amazing that MBBK saw no necessity to prove this assumption. It's even more amazing that MBBK pointed out a prediction that emerges from their thesis, but then didn't bother to check it. So we did so ourselves.

We examined the influence of the variations on their "cooked" list for *War and Peace.* The list is that given in Table 2 of their paper (with the dates they chose). The variations are those presented in tables 5-10 (excepting the 33 variations to the first power that we disqualified in chap. II, Sec 1(A)). The results are as follows:

	r1	r2	r3	r4	Min(r1-r4)	Min(r1-r2)
better	57	58	43	52	52	58
equal	1	4	2	2	9	11
worse	44	33	57	41	41	33
not worse	58	62	45	54	61	69
total	102	95	102	95	102	102

Table 24

Bar-Natan and McKay avowedly "cooked" their list according to r4 (and thus practically determined the results for Min(r1-r4)) with optimization chiefly of the appellations, but also considering the dates and which rabbis to include in the sample. Nevertheless, for completeness' sake we have brought the rest of the rankings. For all of the rankings – there is no indication of optimization. Even if we examine the results for the P-statistics (which should testify to indirect optimization, according to MBBK) we detect no sign of optimization:

	P1	P2	P3	P4	Min(P1-P4)	Min(P1-P2)
better	55	64	38	59	57	66
equal	7	5	11	7	6	10
worse	40	26	53	29	39	26
not worse	62	69	49	66	63	76
total	102	95	102	95	102	102

Table 25

The picture is crystal clear: If MBBK's "study of variations" was correct, and if the variations were chosen (created) without bias - we have here strong proof that Tolstoy intentionally hid codes in "*War and Peace"*. Of course this absolutely contradicts MBBK's claim that finding codes in *"War and Peace"* was just a parody!

2. A Further experiment: Examination of a Sample attributed to Dr Emanuel:

In the previous section we learned how MBBK's variations affect sample that underwent optimization, and we saw that the variations' effect on the result was contrary to that expected by MBBK's thesis.

Now we wish to examine what happens when we apply the variations to a sample that underwent "treatment" that is the opposite of optimization, a sample treated so that its significance should worsen.

(A) In this case too, MBBK supply us with the sample. In chap. 10 of their paper MBBK report several lists of names and appellations prepared by Dr Emanuel, an independent objective expert unilaterally hired by them. MBBK write emphatically and expansively of the experiment that was intended "to mimic" that of WRR. According to our investigations, the sequence of events was as follows (full details of this sorry affair can be found in our paper [8]):

(1) Dr Emanuel was requested to prepare a list of names and appellations for 35 personalities (including 32 personalities of L2) as a substitute for L2, without his seeing L2. We will call this list "list c".

(2) MBBK omitted from this list, without Emanuel's knowledge, two personalities that were in L2. They published the remaining 33 names and appellations, which we will call "list c1", in the name of Emanuel.

(3) Dr Emanuel also prepared the dates for the list.

(4) Thus was created a sample based on the names and appellations of "list c1". We will denote it by EM3(1). This sample was intended for an experiment that MBBK claimed would mimic that of WRR. They claim that EM3(1) was created objectively. Therefore according to their model, applying the variations should provide as "better" cases as "worse" cases. They cannot claim that optimization was caused by their involvement, (see (2) above). See our article [8] where we clearly show that their involvement was intended to deteriorate the results.

(B) Let us apply the "study of variations" for EM3(1). Here are the results for such a study using the P-statistics:

	P1	P2	P3	P4	Min(P1-P4)	Min(P1-P2)
better	5	21	8	17	21	21
equal	20	8	23	7	15	15
worse	77	66	71	71	66	66
not worse	25	29	31	24	36	36
total	102	95	102	95	102	102

Table 26

Here are the results using the r-statistics:

	r1	r2	r3	r4	Min(r1-r4)	Min(r1-r2)
better	17	16	16	17	17	16
equal	14	10	13	8	15	17
worse	71	69	73	70	70	69
not worse	31	26	29	25	32	33
total	102	95	102	95	102	102

Table 27

Once again, a contradiction to MBBK's thesis: This time, a list that was not optimized (on the contrary: this sample was “treated” so that its significance should worsen), is shown by “the study of variations” to be optimized!

3. Two additional experiments:

(A) MBBK chose to present the results of applying their study to L1, through Min(r1-r4). To do this they had to measure the four statistics r1, r2, r3 and r4. Measuring r3 and r4 (by their definition) compelled them to remove a group of appellations from the sample: The standard appellations of the type “Rabbi x”. We will denote the partial sample obtained from this group by RABBI1. The value of the P-statistics for RABBI1 is:

$$P1=6.88\times10^{-4},\quad P2=1.07\times10^{-3}.$$

This group played an important part in the success of L1 in the P-statistics. According to MBBK’s thesis, L1’s success stemmed from direct optimization of the measurement’s parameters, and also from optimization of the data. Therefore, according to their thesis, implementation of the "study of variations" for RABBI1 must indicate clear optimization of RABBI1, optimization aimed at improving the P1 or P2 values of the complete sample of L1.

	P1	P2
better	35	50
equal	14	8
worse	53	37
not worse	49	59
total	102	95

Table 28

However, the results clearly demonstrate that there was no optimization. We emphasize once more, as we explained in chap. III, that the examination of the variations’ effect must be made with the P-statistics that were used for the original experiments. We present the results for the permutation test (r-statistics), only in order to complete the picture:

	r1	r2
better	41	45
equal	9	16
worse	52	34
not worse	50	61
total	102	95

Table 29

Here too there is no sign of optimization.
In conclusion: According to MBBK's thesis we have here clear proof that no optimization was made on the parameters of the first experiment, nor on the data (appellations and dates) relating to the RABBI1 group.

(B) Let us conduct a similar experiment for L2. We will denote by RABBI2 the partial sample obtained from the group of standard appellations of "rabbi x" type in L2. The value of the P-statistics for RABBI2 is:

$$P1=9.28x10^{-3},\quad P2=2.17x10^{-2}.$$

This group played part in the success of L2 in the P-statistics. However, MBBK do not claim here that there was an optimization of parameters—because these were already established in the first experiment. Even the very inclusion of a group of this appellation type was established in the first experiment. Nevertheless, according to MBBK's papers [2][16], at least the two following optimizations were made in this group:

- Concerning the inclusion or omission of rabbis.
- Concerning the dates.

Let's examine the results of variations on RABBI2:

	P1	P2
better	2	39
equal	21	18
worse	79	38
not worse	23	57
total	102	95

Table 30

The difference between the two statistics' results is conspicuous, and it remains even if we move to the r-statistics:

	r1	r2
better	8	40
equal	17	28
worse	77	27
not worse	25	68
total	102	95

Table 31

We will discuss the significance of the contradiction between the results in the next section. Here we will just mention that according to MBBK's thesis, we should prefer the statistics P2 and r2. Therefore, according to them, there is clear evidence that the two above optimizations were never made!

4. Contradictory results:

Let us sum up the picture gained so far from applying the "study of variations" to the various samples. We will use the results presented in the previous sections of this chapter, and the complete results for L1 and L2 presented in chap. III section 1.

(A) We divide the statistics used for the various tests, into two categories according to their results in the "study of variations". If the particular statistic indicates a percentage of "worse" higher than 70% (an arbitrary threshold), we will say that there is an "indication of optimization" for this statistic. If not, we will say that there is "no indication of optimization" for that statistic.

Sample	Indication of Optimization	No Indication of Optimization
L1	P2, P4, r2, r4, Min(r1-r2), Min(r1-r4).	P1,P3,Min(P1-P2),Min(P1-P4), r1, r3.
L2	r2, r4, Min(r1-r2), Min(r1-r4).	P1,P2,P3,P4,Min(P1-P2),Min(P1-P4), r1, r3.
BM Sample in War & Peace	None	P1,P2,P3,P4,Min(P1-P2),Min(P1-P4), r1, r2, r3, r4, Min(r1-r2), Min(r1-r4).
EM3(1)	P1, P4, r2, r3, r4.	P2, P3, Min(P1-P2), Min(P1-P4), r1, Min(r1-r2), Min(r1-r4).
RABBI1	None	P1, P2. r1, r2.
RABBI2	P1, r1.	P2, r2.

Table 32

[Remarks:

(a) We must remember that there are strong dependencies between the variations. Taking such strong dependencies (that deterred even MBBK from quantifying their study's results) into account, we consider 70% or 80% very moderate thresholds. ALL of the data needed for this table is given in this paper, so the reader may check any other threshold.

(b) While summing up the results, as was done in Table 32, we must to clarify what should be done with "ties".

1. According to MBBK's null hypothesis, if WRR were right and there is a code in *Genesis*, the probability for the result to weaken (or to improve) by a variation should be 0.5. According to such a model, one half of the "ties" may be counted with "better" cases, and half with the "worse" cases.
2. According to MBBK's working hypothesis, the result of WRR was obtained merely by "tuning", and we expect the result to weaken when applying a variation. Therefore, a "tie" is against their hypothesis more than for it.
3. One of the two manifest purposes of MBBK's "study of variations" is to check the "robustness" of WRR's result. In this case a "tie" is an evidence *for* "robustness".
4. In a case where we need to quantify a "tie", it seems reasonable in light of points (i)-(iii), that the weight of a "tie" is not the same as that of an "improvement", and it lays somewhere between 0.5 and 1.
5. In the present discussion, where the notion of the summation itself is only to give some rough estimate, we summed up the "ties" with the "improvements" (that is, a weight of 1 for "ties"). For comparison, we present here also Table 32a, where the data are calculated according to weight of 0.5: In this case we put half of the "ties" with the "better" cases and half with the "worse" cases.

Sample	Indication of Optimization	No Indication of Optimization
L1	P2, P3, P4, r2, r3, r4, Min(r1-r2), Min(r1-r4).	P1, Min(P1-P2), Min(P1-P4), r1.
L2	 r2, r4, Min(r1-r2), Min(r1-r4).	P1,P2,P3,P4,Min(P1-P2),Min(P1-P4), r1, r3.
BM Sample in War & Peace	None	P1,P2,P3,P4,Min(P1-P2),Min(P1-P4), r1, r2, r3, r4, Min(r1-r2), Min(r1-r4).
EM3(1)	P1,P2,P3,P4,Min(P1-P2),Min(P1-P4), r1, r2, r3, r4, Min(r1-r2), Min(r1-r4).	None
RABBI1	None	P1, P2. r1, r2.
RABBI2	P1, r1.	P2, r2.

Table 32a

We see that the only significant change is for EM3(1). End of remarks].

The only clear thing in table 32 is that its data is contradictory and cannot possibly be reconciled with the "study of variations".

(1) The results obtained for BM Sample, the sample where optimization was openly done, shows no signs of optimization. The "study of variations" completely collapses here.
(2) The results obtained for EM3(1) show (partially) an "indication of optimization", although according to MBBK no optimization was done (actually we know that it passed a special "treatment" to worsen its significance).
(3) The fact that there are contradictions between the various results for L1, and between the various results for L2, points to the same conclusion. And matters only become worse if one considers the claims of each side of the debate:
(a) If, as MBBK claim, optimization was made in these samples, then:

- For L1: Why does statistic Min(P1-P2) – in reference to which optimization was supposedly made - show no sign of optimization? And why, on the other hand, for P4, which was first defined long after the experiment on L1, there is indication of optimization?
- For L2: Why do most of the statistics reveal no indication of optimization, especially those for which optimization was supposedly made. Meanwhile, only statistic r4 (and the three statistics strongly dependent on it—r2, Min(r1-r2), Min(r1-r4)) indicates optimization, even though the alleged optimization was not done in reference to it?
- And why, according to them, is there no indication of optimization in the RABBI1 sample? And why for RABBI2, which is part of L2, is there no indication of optimization with the statistic r2, which shows optimization in L2?

All this is in addition to the questions we already raised in chap. III.
(b) If, as we claim, there was no optimization in L1 and L2, why do certain statistics appear to show optimization?

(B) To analyze why the "study of variations" gave such strange results, we will first repeat the two main arguments on which it rests.

(1) Optimization of data will appear as optimization of parameters.
(2) The collection of variations was chosen correctly and without bias.

From the data of Table 32 we deduce that it is impossible that both (1) and (2) are correct. Obviously only three possibilities remain:
I. (1) is incorrect.
II. (2) is incorrect.
III. Both (1) and (2) are incorrect.

Examination of assumption I.
Let us examine the possibility that the only reason for the above results is the non-validity of (1). According to this, the picture obtained for the BM Sample and for EM3(1) would be understandable. Perhaps MBBK themselves would like to use this conjecture to explain the picture obtained for RABBI1 and RABBI2.
But this conjecture does not explain why there is such a great contradiction in the results of especially L1 and L2, and why these contradictions are so similar:
(a) In both samples, using statistic Min(P1-P2), for which the alleged optimizations were designed, there is no indication of optimization, while with both samples the main indication of optimization comes from the statistic r4 (and the three statistics strongly dependent on it - r2, Min(r1-r2), Min(r1-r4)).
(b) Strangely, the results for r2 in both samples are very similar, even though the samples are different and built from different word pairs, and even though (1) is incorrect! It is even stranger if one remembers that according to MBBK's thesis, the method of optimization in both samples was different. For L1, the optimization was of the parameters themselves and also of the data and all the details of the experiment, while for L2, all parameters were already established and the alleged optimization concentrated on the data.
(c) The similarity between the results for r4 in the two samples is even more surprising. Especially considering that for L1, the partial group of appellations used to measure r3 and r4 was not at all defined in the original experiment.

Conclusion: The table's results cannot be explained according to this assumption.

Examination of assumption II.
Let's examine the assumption that the only cause of the above results is non-validity of (2). In other words, we must examine whether the picture arising from table 32 is merely the result of defective sampling of the variations. There are two possible reasons for defective sampling:
a. Error: Unintentional choice of variations that are erroneous due to various reasons, dependencies between the variations etc.
b. Tuning: Intentional choice of variations in order to achieve the desired results.

Examination of possibility a:
If the whole failure of the "study of variations" rests solely in unintentional choice of "incorrect" variations, it is difficult to explain:

- Why did this choice "damage" specifically sample L1 and L2, which were singled out as MBBK's target (and also “damage”, as a by product, EM3(1) which contains many “appellation-date” pairs of L2), while it "benefits" the other samples, especially the BM Sample.
- All the objections we raised against assumption I.

Therefore this possibility cannot explain the results.

Examination of possibility b:
According to this possibility, the results of table 32 are the result of "tuning". MBBK deliberately chose variations to reach what they emphasize in their paper - that the statistics they chose for L1 and L2 indicate optimization.

This possibility explains:

- The strong correlation between the statistics (especially the r-statistics) chosen by MBBK and the fact that they indicate optimization for L1 and L2 (as we showed in the first part of chap. III).
- The puzzling questions (a)-(c) we raised against assumption I (which are the same questions raised against "possibility a" in this hypothesis):

(a) MBBK's main efforts in their "tuning" of variations was directed against the result of WRR, that is Min(r1-r4), and that is why statistics like Min(P1-P2) were affected less. As we will show in the next chapter the variations damage results mainly through affecting the permutation test, and therefore they affect the P-statistics far less. In addition, the variations were "tuned" to affect the statistics based on P2 (P4), and not those based on P1 (P3).
(b)-(c) The similarity between the results for r2 and r4 for the two samples L1 and L2 reflects MBBK's naive expectations.

- The results for EM3(1) which contains many "appellation-date" pairs of L2, especially from the group of pairs for which P4 and r4 are defined (more about this in chap. V sec. 2(C)).

This possibility is supported also by the evidence we brought in Chaps. II and III for "tuning" of variations. This possibility also fits in with our claim that there was no optimization of RABBI1, RABBI2, L1 or L2, and therefore:

- Statistics other than those presented by MBBK for L1 and L2 show no indication of optimization (except in case of strong dependence on the MBBK's chosen statistics).
- The statistics chosen by MBBK which are relevant to the samples RABBI1 and RABBI2, that is P2 and r2, show no indication of optimization in both samples.

All that remains is to investigate the results of the BM Sample: Why does it show no indication of optimization? Possibly, the intentional choice of variations to "prove" optimization for L1 and L2, created a defective collection of variations (for example, dependency was set between variations) which created distorted results for the BM Sample.

Conclusion: The results of table 32 can be explained as the outcome of "tuning": Deliberate choice of variations to reach the desired goal.

Examination of assumption III.
Let's examine the assumption that the picture of table 32 comes from non-validity of both (1) and (2). In other words, perhaps these results reflect the failure of MBBK's hypothesis that optimization of data is manifested as optimization of parameters, and

also resulted from defective sampling of variations. According to our discussion of the two previous assumptions, we need only examine the combination of the following two reasons:

a. Optimization of data does not manifest itself as optimization of parameters.
b. "Tuning".

We saw before that reason b is sufficient in itself to explain the results of table 32. Therefore the combination of a and b also suffices to explain it.
Conclusion: assumption III can also explain the results of table 32.

In conclusion:
The data in table 32 can be explained as a (direct and indirect) result of "tuning" of variations to "prove" optimization of L1 and L2.

On the other hand, these data are insufficient to invalidate MBBK's hypothesis that optimization of data manifests itself as optimization of parameters.

In the next chapter we will investigate more thoroughly how MBBK got their results for L1 and L2. We will do this by suggesting an alternative model, and subjecting it to further experiments.

CHAPTER V

AN ALTERNATIVE MODEL: THE EVOLUTION OF THE "STUDY OF VARIATIONS"

In the previous chapter, serious contradictions between the variations' results in a variety of experiments were presented. The data was analyzed, and the conclusion was that the variations were "tuned" intentionally, i.e. the variations were chosen to "prove" optimization in L1 and L2.

In this chapter we go a step further and present a model explaining the experimental results. We will explore the possibility that the process of choosing variations aimed at challenging the only result published by WRR in their *Statistical Science* paper: The result of the permutation test (Min(r1-r4)) for the second list (L2). This is an alternative to MBBK's model.

The "Study of Variations" passed a long and tortuous evolutionary process. The evolution proceeded through two routes: free adding (and removing) of variations, and free choice of what to publicize and what to hide.

In the first part of this chapter we show how the picture created by MBBK gradually moved in a direction which satisfied their ultimate goal: Proving deliberate optimization of the choice of appellations for WRR's second list.

In the other section of this chapter we bring results of additional experiments indicating that MBBK's result is artificial: There is actually no connection between MBBK's thesis of optimization and the result they present for L2.

To save returning to earlier parts of this paper, some things are repeated.

1. The Evolution of MBBK's Data:

In the introduction to their paper, MBBK briefly describe WRR's experiment on L2, and quote its result. Thenceforth, when MBBK use the words "experiment" and

"result" (of WRR) without being specific, their intention is the experiment on L2 and its result. This experiment is the subject of their paper:

"This paper scrutinizes almost every aspect of the alleged result." (Pg. 151)

Explaining the goal of the "Study of Variations" both in the Introduction (Pg. 151-152) and at the beginning of Sec. 7 of their paper, MBBK relates only to the L2 experiment.

Thus their main investigation is directed against L2. Indeed it is necessary to explain why L1 was also included in their study. At the end of Sec. 3 of their paper they make some attempt:

> "WRR's first list of rabbis and their appellations and dates appeared in WRR94 too, but no results are given except some histograms of c(w,w') values. Since WRR have consistently maintained that their experiment with the first list was performed just as properly as their experiment with the second list, we will investigate both." (Pg. 154)

Therefore, we will concentrate here on the evolution of the variations concerning the result of L2.

(A) The True Results:

Let's see what happens if we make the choices **natural** to their thesis:

- We examine the influence of the variations on L2.
- Originally, P1 and P2 were the only statistics used to estimate the success of L2. Therefore any optimization must have been made in relation to P1 or P2, or more likely, in relation to Min(P1-P2). Thus the natural choice is to examine the picture in relation to these values.

These are the results for L2:

	P1	P2	Min(P1-P2)
better	35	38	42
equal	21	6	10
worse	46	51	50
not worse	56	44	52
total	102	95	102

Table 33

There is no indirect evidence here for any optimization! On the contrary: If MBBK's thesis of the "study of variations" is correct, we have clear evidence that there was no optimization!

We claim that all their choices of which results to show us, and all the excuses they invent to justify them, hide this basic fact— as we will see later.

(B) Mutations:

The results in Table 33 are a blatant contradiction to MBBK's much publicized report that WRR's results were almost always worse than the variations.

For example in *Chance* [6]:

> "We reiterate that out of all the cases we looked at, which by now number in the hundreds, WRR's choices were fortunate uncannily often". (Pg. 51)

To understand the cause of their differing report we will bring another excerpt [4]:

> "Wonder of wonders, however, it turns out that almost always (though not quite always) the allegedly blind choices paid off: Just about anything that could have been done differently from how it was actually done would have been detrimental to the list's ranking in the race". (pg. 18)

The end of this quote ("the list's ranking in the race") reveals the source of the difference. They examined the variations not as relating to P1 and P2, which were the sole statistics used to estimate the success of L2, but in relation to the permutation test suggested two years after the supposed optimization. We already discussed their *a posteriori* excuses for this strange step in the second section of chap. III. Here we will just sum up the results they chose to present for L2:

	P4	Min(r1-r4)
better	31	4
equal	7	13
worse	57	85
not worse	38	17
total	95	102

Table 34

MBBK's chief emphasis was on the right column, and they write:

> **"Conclusions.**
> As can be seen from the Appendices, the results are remarkably consistent: only a small fraction of variations made WRR's result stronger and then usually by only a small amount. This trend is most extreme for the permutation test in the second list, the only success measure presented in WRR94." (Pg. 169, emphasis ours)

It should be noted that due to their "mistake" of including 33 additional variations which are actually repetitions of one single variation (of taking the square root, see chap. **II, 1(A)**) — the results from their point of view are even more pronounced: Only 4 improvements out of 135 variations! This value suited their naive expectations quoted earlier from *Chance*, or in their latest paper (especially on pg. 169).
(Remember: When they wrote that "the results are remarkably consistent", they mean the results that they choose to show us, and not, for example, the true ones given in Table 33).

In light of the tremendous difference between table 33 and table 34, one may well ask: What kind of mutation caused the balanced results of Table 33, to transmute to the extreme one sided results of Table 34? To answer this let's trace the evolution itself.

(C) Seeing the Evolution in Action:

First let's list the known stages of the "study of variations".

1. Bar-Hillel's study of Jan. '97 examined how many times WRR benefited from 13 "choices" available in their experiments. Her study was done using r1 and r2, and

two other statistics that WRR never used at all. Prof. Bar-Hillel announced [21] a result of 13 to 0 against WRR. Prof. Aumann's piercing criticism of her work can be seen here [9]. On 5 Jan. '99, during a discussion after our lecture at the "Center for Rationality and Interactive Decision Theory" in the Hebrew University of Jerusalem, Bar-Hillel announced that she had thrown her work into the "waste basket".

2. McKay's first report, Feb. '97. McKay checked the "robustness" of L2 for only four cases using r1 and r2. He also used three additional statistics that WRR had never used.

3. McKay's second report, April '97. This is actually the first substantial study on this issue. Here McKay asks, "Where is the low probability?" and attempts to show, through testing of variations, that the parameters of the experiment leave much room for maneuver. Here the rationale for the use of variations is that: If many variations give worse results, the conclusion is that the result of the original experiment (0.00002) was due to deliberate choices made in the experiment. He examined 20 groups of variations, most of which had more than one sampling point. He presents the results for P1, P2, r1 and r2 for both Rabbis lists.

4. BBM's (Bar-Hillel, Bar-Natan & McKay) article in *Chance*, spring '98. Here we find a selective presentation of examples of variations, some of them presented for the first time. All of them are measured for r2 alone, with no mention that the measurements were done using other statistics as well; no reason for this is given. The emphasis is on L2, and for most variations results are also given for L1.

5. MBBK's latest paper in *Stat. Sc.* May '99. Many variations are presented for the first time (but not so many considering the hundreds of variations which MBBK admittedly [6] checked). This time MBBK chose to present the results for the following statistics: P2 and min(r1-r4) for L1, and P4 and min(r1-r4) for L2.

Thus we find a process of evolution both for the variations themselves and for the choice of which results to present.

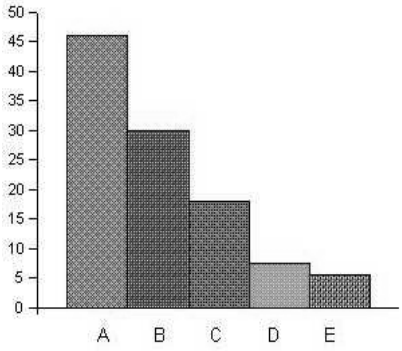
In order to understand how MBBK got the extreme results for min(r1-r4) in Table 34, we will now trace what happened in the following stages to the results of the variations for L2 in the r-statistics.

Let us denote:
A = Base of comparison: The true results for min(P1-P2) (Table 33).
B = The presentation method of stage 3 (r1 and r2). The variations are the same as those whose results appear in McKay's report and which are also found in tables 5-10 of MBBK's latest paper.
C = The presentation method of stage 3 (r1 and r2). The variations are all those variations found in tables 5-10 of their latest paper.
D = The presentation method of stage 4 (r2). The variations are all those found in tables 5-10 of their latest paper.
E = The presentation method of stage 5 [min(r1-r4)]. The variations are those found in tables 5-10 of their latest paper.

We show here what happened in the various stages to the percentage of variations where the results improved in the permutation test. (The results are according to what MBBK wished to show us: that is, according to their mistaken null hypothesis, and when the "tied" cases are equally divided between the "improvement" cases and "worse" cases):

Stage	Improvement Percentage
A	46.1
B	30.0
C	17.9
D	7.4
E	5.5

Table 35



A graph plotting the drop of the improvement percentage over the years.

Now that we have followed the evolutionary process itself, we wish to clarify how the mutation we saw in Table 34 was created. We will do this in the following section.

2. Where is the "optimization"?

Through a process of elimination, let's investigate the exact data in L2 which supposedly underwent optimization.

(A) When MBBK's paper analyses "tuning" of data in L2, it lists three components of the freedom, which may enable the hypothetical "tuning":

a. Freedom in "choice of rabbis" (pg. 155 in their paper).
b. Freedom in "choice of dates" (pg. 155 ibid).
c. Freedom in "choice of appellations" (pg. 156 ibid).

(1) Of these three components, MBBK considers *component c* the major one: They stress this throughout their paper. For example, at the beginning of Sec. 7:

> "In the previous sections we discussed some of the choices that were available to WRR when they did their experiment, and showed that the freedom provided just in the selection of appellations is sufficient to explain the strong result in WRR94." (Pg. 157)

In other words, they claim that *component c* suffices to explain L2's success.

On the other hand, it is easy to see that no use was made of components a and b to improve WRR's success: Because establishing the rabbis' names according to MBBK's criteria, together with the dates supplied by their expert, gives a list with ***greater success.***

Let us specify: Establishing the rabbis for L2 according to MBBK's criteria [16,17], and using dates according to their expert, gives a **more successful** list [using the measure for success used at that stage: min(P1-P2)]. This **improvement** is by a

factor of 3.4. (Note: Even if we accept their dubious argument to omit R. David Ganz, we still have an **improvement** by the factor of 1.8. And even if we use the very latest criteria that MBBK just invented for including rabbis ([8]: chap. I, 2(B)(3)), we still have an **improvement,** albeit by a lower factor).

Thus MBBK themselves, through their choice of criteria for inclusion of rabbis in L2, and through their "correction" of dates, help to prove that ***our choices were to our detriment****, thus proving that they were made without bias.*

(2) In conclusion we are now left with *component* c_1 i.e. with the possibility that the optimization of L2 was through the appellations. Do they claim this was possible *with all* the appellations?

It turns out that even according to their claims (see chap. III part 2(b)) no optimization was conducted with the group of standard appellations of type "Rabbi X". We will refer to the group of the remaining appellations as L'2. So according to MBBK optimization was done specifically on the appellations of L'2, and their proof is the extreme result obtained for min(r1-r4). The result for min(r1-r4) in their "study of variations" comes solely from the results for statistic r4, and this statistic, by definition, measures only the appellations in L'2 (the statistics relating to L'2 are limited to: P3, P4, r3, r4).

We have to investigate: Exactly where is the "optimization" in L'2?

(B) First let us copy the results for L'2 from tables 20-22:

	P3	P4	r3	r4
better	52	31	53	4
equal	14	7	11	6
worse	36	57	38	85
not worse	66	38	64	10
total	102	95	102	95

Table 36

It is clear from this table that only r4 indicates any optimization. MBBK raised various arguments to prefer r4 over P4. In chap. III (the section of "pretexts") we already dealt with all their *a posteriori* justifications for this. One of their arguments was that the improvements using P4 derived from a certain "tendency" which we called the "charisma" of the appellations. Therefore, they concluded that the right statistic is r4, which cancels this "charisma". See there how we refute their conclusion.

At this stage we wish to see whether this is the true reason for the great difference in results between r4 and P4.

(1) <u>There is a simple way to do this</u>: We calculate the c-values of the "appellation-date" pairs, when the *appellations* are taken only as ELSs. This way the 'charismatic" effect of the appellations is nullified [20]. We do it with the variations, and use P4. For comparison, we do exactly the same thing while calculating the c-values of the "appellation-date" pairs, this time when the *dates* are taken only as ELSs.

	Dates only as ELSs	Regular calculation	Appellations only as ELSs
better	26	31	29
equal	7	7	7
worse	60	57	57
not worse	33	38	36
total	93	95	93

Table 37

[For the left and right rows there are only 93 variations, because two of the variations are impossible using “appellations only as ELSs” and “dates only as ELSs”]

We see that the results are similar with no significant differences between them. On the other hand, the result using r4 is totally different. This proves that the result for r4 results from some other cause, which we will try to clarify later. Thus we have experimentally disproved MBBK’s argument. In the rest of their arguments, MBBK fail to give even *one* valid reason to prefer r4 over P4 (see chap. III, the section of "pretexts").

(2) We would like to point out that the big difference between the variations’ results for P4 and r4 seems very exceptional. The following data will demonstrate this. Let us define:
Imp(P4)=The number of improvements for P4,
Imp(r4)=The number of improvements for r4,
And let Q=Imp(P4)/Imp(r4). Than we get for the various samples (see chap. IV):

Sample	Imp(P4)	Imp(r4)	Q
L1	17	6	2.83
L2	31	4	7.75
BM Sample in War & Peace	59	51	1.16
EM3(1)	17	17	1.00
RABBI1	51	45	1.13
RABBI2	39	40	0.98

Table 38

[P2 and r2 were taken for RABBI1 and RABBI2 since P4 and r4 are not defined for them].
We think that the exceptional result for L2 is a result of the “tuning” done with the variations, whereas the smaller value for L1 is a byproduct of it.

In Conclusion:
In our opinion, the correct method is to use the P-statistics for the "study of variations". Therefore, for L’2 the right checking method is with P3 and P4. Only to see further how the variations were "tuned", will we also examine the results for the r-statistics (r3 and r4).

(C) In Sec. 10 of their paper, MBBK report on several lists of names and appellations prepared for them by Dr Simcha Emanuel. We publicized a special paper [8] dealing with these lists. One of these lists, which we called "list c", was intended “to mimic” L2. We will now use this list to conduct an experiment.

Let's take from "list c" the appellations of the 32 rabbis of L2, which have 5-8 letters (this was done in the original experiment). We will denote this group by EM3. We will prepare two groups of appellations:

- Group *A* = the intersection of EM3 with L'2.
- Group *B* = L'2−*A*.

A includes those names and appellations that Dr Emanuel chose as well, minus a few appellations [there are altogether 6 such appellations (or an alternative spelling of the same appellation), one of which has no ELSs in Genesis and therefore is not relevant to the present discussion]. Therefore, it would be reasonable for the behavior of *A* and EM3 to be similar under "the study of variations"

On the other hand, *B* includes the rest of L'2's appellations: Exactly those chosen by Prof. Havlin and not by Dr Emanuel. So if there was any optimization, it must have been on the appellations of *B*.

Let us check groups *A*, *B* and EM3, using the original statistics for L'2: P3 and P4.

	EM3		A		B	
	P3	P4	P3	P4	P3	P4
better	9	12	9	16	60	41
equal	26	11	30	8	16	10
worse	67	72	63	71	26	44
not worse	35	23	39	24	76	51
total	102	95	102	95	102	95

Table 39

Summary of results:

- According to MBBK's thesis, EM3, which underwent no optimization of the appellations, should not look like an optimum under variations. But the exact opposite occurred: Using P4, EM3 looks like an optimum compared to the corresponding results for L'2.
- The results for *A* are similar, as expected, to the results for EM3. Again, it is strange that *A* exhibits an optimization compared with L'2.
- On the other hand, according to that same thesis of MBBK: *B*, which contains exactly those appellations that underwent the alleged "optimization", should exhibit sharp optimum under the variations. But instead the opposite occurs. There is absolutely no optimum for *B*!

(D) Let's check once more, this time according to the r-statistics:

	EM3		A		B	
	r3	r4	r3	r4	r3	r4
better	13	14	15	11	72	22
equal	9	7	13	10	9	10
worse	80	74	74	74	21	63
not worse	22	21	28	21	81	32
total	102	95	102	95	102	95

Table 40

Summary of results:

- The results for EM3 and *A* remained essentially unchanged.
- On the other hand, for *B* the result of r4 is very different than the result of P4: The number of "worse" results rose by 43%.
- Nevertheless, the results continue to be surprising:

- Specifically for *B* which underwent the alleged "optimization", the number of improvements in r4 is *greater* than that for EM3, the group which was supposedly free of optimization, and *double* of those for *A*.
- Therefore, it also happens that for *B*, the number of improvements for r4 is 22, compared to the *four* improvements listed in tables 5-10 of MBBK. This result is surprising according to the model of MBBK: L'2 is the union of *A* and *B*, therefore it includes the group of appellations which underwent no optimization, *A*. We would expect that this "inert" element would contribute towards a balance between improvements and "worse" cases. But the opposite happens: The number of improvements for *B*, where the optimization was supposedly concentrated, is 5.5 times greater than for L'2.

(E) All the results obtained in the previous paragraphs are an absolute contradiction to MBBK's thesis: There is absolutely no connection between the variations' results and "optimization". But this is nothing new: We already proved in the previous chapter that the results of the "study of variations" are really the result of "tuning" of variations. Here too we see the results of this "tuning": MBBK "tuned" their variations to reach a minimum of improvements for r4 in L2 [and thus they reached a minimum of improvements for min(r1-r4)], But they did not "tune" them with respect to EM3, *A* or *B*. Thus this defective "tuned" collection of variations leads to strange results when applied to EM3, *A* or *B* (as was explained at the end of chap. IV).

But there is still room for further scrutiny of the results of the permutation test for EM3, *A* and *B*. The samples based on EM3, *A* and *B* have many cases where a rabbi has no appellation or date. In such cases there is no contribution of "appellation- date" pairs to the sample itself, but there is an indirect influence through the permutation test. MBBK claim [16] that this causes "random noise" and therefore they removed the data that is not involved directly in the "appellation-date" pairs of the sample itself.

For *A* and *B* this issue reaches an extreme because only less than a half of the rabbis have at least one "appellation-date" pair. We checked what influence such an extreme situation has on the results, by removing the data not involved directly in "appellation-date" pairs in the sample itself.

The results of the permutation test now look like this:

	EM3		A		B	
	r3	r4	r3	r4	r3	r4
better	18	15	17	14	76	37
equal	9	9	14	8	10	8
worse	75	71	71	73	16	51
not worse	27	24	31	22	86	45
total	102	95	102	95	102	95

Table 41

We see from this table that:

- There is a clear change with r4 for *B*: There is a dramatic rise (68%) in the number of improvements, and the results are almost as balanced as for P4.
- There is a more moderate rise (27%) in the number of improvements for *A* in r4. Now, the results are close to those for P4.

Thus the last traces of "optimization" vanish from *B*, the group that underwent the alleged "optimization".

In conclusion:

(1) The summary of the experiments in chap. IV (table 32), together with the experiments in this chapter, are clearly incompatible with the possibility that the hypothesis of MBBK is correct, and that the results of their study is due to optimization of data by WRR.
(2) On the other hand we can explain these results assuming that there was "tuning" of the variations. This explanation is supported by evidence for "tuning" brought in previous chapters.
(3) We have outlined the evolution of how the variations were both created and presented, with a clear attempt to improve MBBK's desired results.
(4) We tried to trace the source of the apparent "optimization" exhibited by MBBK. To do this we used the list of names and appellations of Dr Emanuel (an expert engaged by MBBK) as a database. It transpired that:
    (i) Specifically the names and appellations chosen by Emanuel demonstrated "optimization".
    (ii) On the other hand, the names and appellations chosen by Havlin and not by Emanuel showed no signs of "optimization".
(5) Examination using the r-statistics which MBBK preferred showed "optimization" also in the names and appellations chosen by Havlin and not by Emanuel (but not to the extend of the "optimization" shown by the names and appellations chosen by Emanuel). But in the end we saw that this "discovery" was due to a certain feature of the permutation test. When the 'noise' was removed, the "optimization" also disappeared.
(6) All of this arouses suspicion that the result of MBBK's study for r4 is no more than an anomaly caused by certain feature(s) of the permutation test, with no connection to the existence or otherwise of optimization. If this is correct, it turns that MBBK made the "tuning" of the variations so amateurishly, that the results of the variations depend strongly on some features of the permutation test, and nothing more!

APPENDIX

For Chapter I:

1. For paragraph 1:

In our first preprint ('86) we emphasized the importance of two elements in the geometrical convergence between two ELSs: Each ELS must be "concentrated" on the two dimensional table (or cylinder). In other words they must have a "small localization parameter" (small *f*), and they should be close to one another (small *l*).

See pages 8-9 and 29-30 there. But MBBK ignored this in a sizable number of the variations of Table 5.

2. For paragraph 9:
Let's list the various statistics with which one can measure the variations' results.

(a) First, we will list the possibilities for P-statistics. The default choice is min(P1-P2), and we already pointed out in this paper (at the beginning of chap. III) that it is strange that MBBK ignored this natural choice.
Besides this we have the 4 known statistics: P1, P2, P3, P4.
MBBK also used min(r1-r4). By the same token they could have also used the corresponding statistic, min(P1-P4).
Altogether we have 6 statistics.

(b) In statistics r we have corresponding to them 6 statistics: r1, r2, r3, r4, min(r1-r2), min(r1-r4).

(c) Besides this, Prof. Bar-Hillel used, to test the variations, two statistics which WRR never used, as Prof. Aumann pointed out in his letter to her [9]

(d) Besides this, to test "robustness" (a test which is one of the declared aims of the "study of variations") in his first report [13], McKay used three more statistics which WRR never used.

Therefore, even if we suffice with what is known to us, we will reach 17 possible statistics for each sample. And because MBBK were not particular to choose the same statistics for the same two samples we have:
$N=2^{34}$ possibilities, which is more than 17,000,000,000 possibilities. From this vast number MBBK chose four specific statistics: two for the first sample and two for the second sample.

We do not assert that *all* the possible combinations are equally reasonable. It is quite hard to know how many reasonable stories could MBBK invent in order to justify possible choices of combinations. But MBBK tell us [22] that they have tremendous ability to create such stories, and for them the space of stories is vast and quite unlimited.

ACKNOWLEDGMENTS

We wish to express special gratitude to Dr Shalom Srebrenik for helpful discussions and valuable suggestions. We used software of Yoav Rosenberg and Ya'akov Rosenberg for our experiments, and we thank them.

BIBLIOGRAPHY

1. Witztum, D., Rips, E. and Rosenberg, Y. (1994). Equidistant letter sequences in the Book of Genesis. Statist. Sci. 9 No. 3 429-438. Available at: http://www.torahcode.co.il/pdf_files/pub/wrr.pdf.
2. McKay, B. D., Bar-Natan, D., Bar-Hillel, M. and Kalai, G. (1999). Solving the Bible Code puzzle. Statist. Sci. 14 No. 2 150-173.
3. Witztum, D. (2000). Of Science and Parody: A Complete Refutation of MBBK's Central Claim. Available at: http://www.torahcode.co.il/english/pdf_files/parody1e.pdf
4. Bar-Hillel, M., Bar-Natan, D. and McKay, B. D. (1998). Torah codes: puzzle and solution. Chance 11 No. 2 13-19.

5. McKay, B. D. (April 1997). Equidistant letter sequences in Genesis – A Report (draft).
6. Bar-Hillel, M., Bar-Natan, D. and McKay, B. D. (1998). Reply, Chance 11 No. 4 50-51.
7. Witztum, D. (1999). Concerning the statistical test that was published in our paper in Statistical Science, Part B. Available at: http://www.torahcode.co.il/english/pdf_files/persi4e.pdf.
8. Witztum, D. (2000). New statistical evidence for a genuine code in Genesis. Available at: http://www.torahcode.co.il/english/pdf_files/emanu1e.pdf.
9. Aumann, R. J. (1997). A letter to Maya Bar-Hillel, dated 17 Jan. Available at: http://www.torahcode.co.il/auman_to_maya.txt
10. Witztum, D. (2000). Concerning the Choices of Dates for WRR's Rabbis Samples, Available at: http://www.torahcode.co.il/english/pdf_files/date1e.pdf. Part B. In preparation.
11. Gans, H. J. (2000). A Primer on the Torah Codes Controversy for Laymen. Part A: http://www.torahcode.co.il/pdf_files/oppose/primer-final-1.pdf, Part B: http://www.torahcode.co.il/pdf_files/oppose/primer-final-2.pdf, Part C: http://www.torahcode.co.il/pdf_files/oppose/primer-final-3b.pdf.
12. Private communication.
13. McKay, B. D. (Feb. 1997). Equidistant letter sequences in Genesis – A Report (draft).
14. Witztum, D., Rips, E. (1998). Reply: Choice of Choices. Chance 11 No. 4 48-49. http://www.torahcode.co.il/english/ chance.htm
15. Witztum, D., Rips, E. and Rosenberg, Y. (1986). Equidistant letter sequences in the Book of Genesis. Preprint.
16. Bar-Natan, D. and McKay, B. D. (1997). Equidistant letter sequences in Tolstoy's "War and Peace" (draft). http://cs.anu.edu.au/ ~ bdm/dilugim /WNP/draft.
    Bar-Natan, D. and McKay, B. D. (1999). Equidistant letter sequences in Tolstoy's "War and Peace". http://cs.anu.edu.au/ ~ bdm/dilugim /WNP.
17. Bar-Hillel, M., Bar-Natan, D. (1996). A letter to R. Aumann, dated 27 Nov., question no. 5. Found in: Document 2 (1997). Bar-Hillel and Bar-Natan inquire – Witztum and Rips respond. Available at: http://www.torahcode.co.il/english/pdf_files/docum2e.pdf
18. Kalai, G. (1997). A letter to R. Aumann, dated 11 Nov 97.
19. Kalai, G. (1997). Letters to R. Aumann, dated 11 and 12 Nov 97.
20. Witztum, D. (1998). Concerning the "REMEZ" in Equidistant Letter Sequences (ELS's). BDD, Journal of Torah and Scholarship, Bar-Ilan University Press, No. 7, , Summer 1998 [in Hebrew]. Available at: http://www.torahcode.co.il/pdf_files/pub/bdd.pdf.
    Witztum, D., Beremez, Y., (1998). The "Famous Rabbis" Sample: A new measurement. http://www.torahcode.co.il/english/pdf_files/new2e.pdf.
21. Bar-Hillel, M. (1997). A lecture at the meeting of the Center for Rationality and Interactive Decision Theory at Zarka Ma'in.
22. Bar-Natan, D., McKay, B. D. and Sternberg, S. (1998). On the Witztum-Rips-Rosenberg sample of nations, Section 3.4. http://cs.anu.edu.au/ ~ bdm/dilugim /Nations.