Bs"d, Shevat 5762 (Jan. 02)

McKay's Reconsideration of his Hypothesis

Following our Criticism of his *"STUDY OF VARIATIONS"*

Doron Witztum

Summary

McKay continues his attempts to distract the public from the simple fact that our series of articles has successfully proved that MBBK's paper in *Statistical Science* (like McKay's other publicized papers concerning the Torah Codes)

- Has no scientific foundation.
- Has falsely "cooked" results.

(See the synopsis "A Review of the Attempts to Invalidate the Torah Codes" and its links to the full articles).

McKay particularly dislikes our paper "MBBK'S 'Study of Variations'" [1], since it proves "that MBBK's case is fatally defective, indeed that their results merely reflect on the choices made in designing their 'study of variations', collecting the data and presenting the results".

McKay has no reply to these articles. Instead he selectively attacks marginal issues or "straw men" of his own creation. In his article [2], "The Analysis of Variations – a reply to Doron Witztum," written in response to our article [1], he repeatedly puts words into our mouths which we never said, and gives selective examples which do not represent the general case. We cannot waste our time on this kind of debate. Instead, we will just give a few comments on this part of [2]. We are sure that the discerning reader who examines our original article [1], will easily find that McKay twisted our arguments out of all recognition and that his claims lack any substance.

Actually, McKay himself recognizes that his pretexts are not convincing. His inability to cover up the failure of his "Study of Variations" forced him to announce **a reconsideration of his hypothesis**. But a close examination of this substitute to his hypothesis shows that it is nothing more than a failed attempt to lay a smokescreen.

I. McKay's claims concerning M1, M2 and M3

McKay argues that our "success measure" was "tuned" in a multi-stage process. He writes:

> "In particular, we want to reconsider whether the success measure was itself tuned. …Note that there are at least three steps in the process:
>
> 1. WRR used some success measure M1 for their first list.
> 2. WRR used some success measure M2 for their second list.
> 3. WRR distributed a program which implements a success measure M3.

> The standard history is that, apart from addition of the permutation test at the second step, all these measures are the same. However, we now know that they are all different."

McKay proceeds to explain in detail:

A. McKay's claim that there is a difference between M1 and M2:

> (1) "We **know** that M1 and M2 are different because the preprints which describe them have differing mathematical descriptions." (Emphasis mine)

For more details he refers the reader to an appendix titled "Early changes in WRR's success measure." At the beginning of this appendix he announces:

> (2) "Here are two examples where the success measure presented by WRR with their first list of rabbis differs mathematically from the success measure presented by WRR with their second list of rabbis".

Our response:

Both McKay's verbosity and the photocopies provided of the first preprint fail to provide **any proof** to bolster his contentions. Let's examine each of the two examples mentioned in his appendix.

Example 1:
His first example pertains to the method of choosing a set of perturbations of the ELSs (Equidistant Letter Sequences). That is, the method by which the set of almost equidistant letter sequences was defined. As is well known [3], the success measure of the convergence between two words appearing as Equidistant Letter Sequences (ELSs), is calculated by comparison to the convergence of the same words appearing as "Perturbed" Letter Sequences (PLSs). McKay's argues that the method of producing the set of perturbations used in the second-list experiment, differed from that used for the first-list experiment.

But:
1. Contrary to McKay's claim (denoted above as (2)), comparing our first preprint [4] (where the first-list experiment was publicized for the first time) to the second preprint [5] (where the second-list experiment was publicized for the first time) proves that **his words are based on thin air.** Indeed, he troubled himself to show the reader a photocopy of the first preprint, but never bothered to bring the second preprint as well. Why not? Because the description in the second preprint **accords** perfectly with that of the first one.
2. McKay himself writes in his appendix that the procedure described in the first preprint could be a mistake, and that the PLSs may have been defined in a different way:
   > "…there is reason to believe it was not actually used to compute the distances that are presented in the preprint"

This is an absolute understatement. McKay **knows very well** that in (what he calls) M1, M2 and M3, there was **only one** method to create the set of perturbations. This can be proved easily, because the perturbation production methods leave prominent "fingerprints."

Let us explain: The evaluation of convergences between the ELSs of the expressions *w* and *w'* is made by comparing them to the convergences between the PLSs of the same expressions. A "contest" is set between the ELSs and a set 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 *v/m* where
v = the ranking of the convergence between the ELSs, and
m = the number of all the competitors for which *w* and *w'* appears as PLSs. (This number includes the ELS competitor.)
For example, *m*=125 means that *w* and *w'* appeared as ELSs and also as PLSs in all the 124 perturbations of the set. But when there are perturbations where *w* and *w'* do not appear as PLSs, then *m*<125. In such cases, t**he value of *m* is typical of the method used to create the set of perturbations.**
Comparing the detailed results of *c(w, w')* in the two preprints to the results of (what McKay calls) M3, demonstrates that **we always used the same method to create the set of perturbations.**

For example: The number of perturbations for the expression "Rabbi Yehudah" is 56 (including the ELSs). Let's examine the results for (what McKay calls) M1, M2 and M3.

M1: In the first preprint there are a number of different results for "Rabbi Yehudah." For example, 8/56 for the pair "Rabbi Yehudah" – "18 Elul."
M2: In the second preprint there are a number of results for "Rabbi Yehudah." For example, 2/56 for the pair "Rabbi Yehudah" – "5 Cheshvan."
M3: In this program we get the result of 2/56, likewise, for the pair "Rabbi Yehudah" – "5 Cheshvan."

On the other hand, using the method described in the preprint for creating perturbation sets, 1/55 is the result for the pair "Rabbi Yehudah" – "5 Cheshvan," and 6/55 for the pair "Rabbi Yehuda" – "18 Ellul."

Note: Had the first experiment used the method for creating perturbations described in the preprint, **its overall result would have been about ten times better.**

Example 2:
McKay's second example pertains to calculating D(*w*), which is the upper bound for the skip in searching for a specific word, *w*, as ELSs (or PLSs). In all our publications we calculated D(*w*) so that in the range of skips $2 \leq |d| \leq D(w)$, the word *w* is expected to randomly appear 10 times on average, as ELSs (or as PLSs).

But:
Here too, despite McKay's assertion in (2) above, that he detected a mathematical difference in calculating D(*w*) between the first and second preprints, not an iota of proof is brought to back it up. Although he presented a photocopy of the first preprint, he failed to present a corresponding photocopy of the second preprint. Why? - Because the second preprint **corresponded** perfectly to the first preprint. McKay himself writes:

> "In the second preprint of WRR, where the second list of rabbis was first presented, this part of the algorithm is described using English prose that can plausibly be read either way."

Was he hoping that no one would notice this sentence?

In Conclusion: McKay's claims are absolutely baseless. His opening assertions (denoted by (2) above) and the examples themselves are totally contradictory. One can hardly escape the conclusion that McKay hoped the reader would swallow his "headlines" without checking the details.

[*Note:* 7 Nissan, 20 March: McKay recently altered his Appendix because of our criticism. Concerning this, see later in our Appendix.]

B. McKay's claim that there is a difference between M2 and M3:

McKay writes:

> "We know that M2 and M3 are different because the program distributed by WRR does not give the same distances between word pairs as are listed in WRR's preprints. Witztum has admitted that there was an earlier program that gave different values but is unable to give it to us. Some of the changes might have been strictly error corrections, but since WRR's later programs still contain errors we don't know whether error correction was performed in a blind fashion (i.e., without regard to whether the result improved)."

Our response:

First of all, the facts:

1. The original results published in the second preprint [5] in 1988 were produced by a program which we will call PROG1.
2. To enhance efficiency, we moved from a VAX/VMS computer to a PC, and from PASCAL to C. This newly recreated program is called ELS1.
3. At the end of 1991, the permutation test was conducted using ELS1, and its results were published in our article in *Statistical Science* [3].

According to Mckay's new hypothesis, ELS1 (which he calls M3) is a result of optimizing PROG1 (which he calls M2), and specially created to give improved results in the permutation test.

But this is nonsense.

It is true that the values of *c(w, w')* obtained by ELS1 are not exactly the same as those obtained by PROG1: In some cases there is a difference between them. But, the differences are usually small and equally distributed in both directions, and the overall outcome is quite the same. Furthermore, the move from PROG1 to ELS1 made the results **worse** for the base-statistics P1, P2, P3, P4:

- In the second preprint, the values of the P-statistics for the second list according to PROG1 were:

$$P1=6.15\ SIGMA,\ P2=1.15X10^{-9},\ P3=5.52\ SIGMA,\ P4=7.20X10^{-9}.$$

- The second list's P-statistic values used for the permutation test were calculated according to ELS1 and they came out **inferior** to the previous results:

$$P1=5.95\ SIGMA,\ P2=1.5X10^{-9},\ P3=5.30\ SIGMA,\ P4=7.71X10^{-9}.$$

Thus the trend of these differences is not to optimize the results but to make them worse!

McKay knows this simple fact. But he insists on claiming “Perhaps." He insists that perhaps, even though the trend shows the opposite, perhaps, nevertheless, there is a craftily hidden optimization, which, he admits, he cannot prove and certainly cannot measure (some new kind of *phlogiston*).

> This reminds me of a similar amusing theory developed by Gil Kalai, who was McKay's co-author in several articles (including the one publicized in *statistical Science*). His theory tried to explain WRR's success by claiming “perhaps”. This theory said that perhaps there were typing errors when WRR’s data was fed into the computer, and perhaps, psychologically, WRR only caught those mistakes which led to “bad” results and not those which led to “good” results. Perhaps this was the source of their overall good result.
>
> He entertained this theory for quite a while. I heard about it from people with whom he had discussed it. One of them asked me what I had to say about it. I asked him if he knew how many “good” results there had been in my experiment. He had no idea. When I told him, about sixty, he laughed. Suddenly he realized that by merely typing about sixty pairs of expressions (which had the “good” results) correctly and working out the results with our freely available program, it could easily be verified whether it was typing mistakes which had caused our “good” results or not.

Now McKay comes along with a new "perhaps" theory, taking advantage of the fact that the PROG1 program became lost. Here too the facts can be checked. The procedure described in the first preprint (excluding the perturbation method, which as we proved in Sec. A, example 1, was exactly the same as in the final article) can be given to a non-biased programmer to produce a PROG'1 and see how it performs in the permutation test. I expect that the results would be much the same as those of ELS1.

McKay himself independently recreated a program similar to ELS1, guided by WRR's final article, and produced very similar results. He also prepared hundreds of programs for his "Study of Variations." Is it conceivable that McKay didn't bother to recreate a program based on the first preprint?

Why did he instead create a web of "maybes," carefully leaving himself an escape hatch by concluding:

> “Nevertheless, the degree to which this explanation is significant is impossible to determine.”

II. McKay’s reconsideration of his hypothesis

McKay was forced to announce a reconsideration of his hypothesis. Why?
The course of events leading to this retreat was as follows:

1. MBBK claimed in their article [6] that they built a "tester" (the “Study of Variations") to examine whether WRR's experiment was "cooked." However, MBBK admitted that the experiment's parameters were not "cooked;" therefore they claimed that the data was "cooked," and that this would manifest itself in their "tester" as if the parameters were "cooked."
2. In our article [1] we proved that MBBK's "tester" is flawed and faked.

In his "reconsideration," McKay produces a new argument in the desperate hope that **this** may save his "tester" from disgrace. He now argues that the "success measure" used in WRR's original experiment was "cooked."

But McKay does not explain why this theory should prove his "tester" any less flawed and faked than before.

Moreover, we already proved in Chap. I that this new claim is merely another useless attempt to lay a smokescreen.

III A rope of sand

McKay has a habit of putting words into our mouth that we never dreamt of, and presenting selective examples which fail to give the true picture. We cannot waste our time on this kind of debate. Instead, we will just give a few comments on this part of [2]. We are sure that the discerning reader who examines our original article [1], will easily find that McKay distorted our arguments and that his claims are worthless.

A. In our article [1] we demonstrated the basic flaws in McKay *et al*'s work:

> "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." (Chap. I, Sec. 6)

We added further criticism:

> **"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!?" (ibid, Sec. 10)

In particular, McKay *et al* often claim that the "vast majority" of (or even "almost all") the results indicate that the results worsen through variation (and this supports their hypothesis), or they stress how few results indicate any improvement. In other words they make extensive use of "raw counts."

In his reply [2] McKay does not deny the existence of basic flaws like these in his "study of variations". He even continues using these methods. For example in his summary he writes:

> "What we found was that, in the great majority of cases, changing a parameter of WRR's experimental method made their result weaker."

This claim is based on "raw counts."
McKay has the philosophy that the best defense is attack. Therefore, instead of **apologizing**, he tries to turn the tables on us and complains that our disproof is based on "raw counts" and that we are trying to create a "psychological impact."

This is laughable.

It is McKay *et al* who rely on psychological impact. We made it clear that relying on raw counts lacks any scientific basis (see above quote) and that McKay *et al*'s work has no scientific value. We only used "raw counts" to prove that **even according to their method of proof,** the impression they create through "raw counts" is false and deceiving.

B. In [1] we showed that McKay *et al* "cooked" the "Study of Variations" by selecting the variations best suited to their thesis. McKay replied (in the section, "Answer to Witzum's claim 1"):

> "The refutation of Witztum's first claim is that he did not manage to identify a single variation which tells a contrary story and should have been presented but was not".

There is not a grain of truth in this reply. Our article [1] is replete with variations which indicate the opposite of McKay's thesis. Therefore he is forced to invent excuses why he didn't include them. In addition he:

- Puts words into our mouth which we never said, and then finds fictional contradictions between them.
- Responds to only one of our many proofs that he "cooked" his variations.

See our article [1] to examine our original arguments and our many powerful proofs that McKay *et al* indeed "cooked" their "Study of Variations."

McKay seems to have had difficulty identifying variations "which tell a contrary story." He obviously did not try too hard. For instance, his reply [2] does mention **one new variation**: The method of calculating D(*w*) [which is the upper bound for the skip in searching for a specific word, *w*, as ELSs (or PLSs)] according to the first (and second) preprint, and not according to our final article in *Statistical Science.* See above: Chap. I, Sec. A, example 2.
The results of this variation are given below:

For the P-statistics:

	P1	P2	P3	P4	Min(P1-P4)	Min(P1-P2)
L1	1.0	0.6	1.0	0.8	0.7	0.7
L2	0.1	0.6	0.1	0.6	0.6	0.6

(We used the following notation: L1=the first list, L2=the second list.)

For the r-statistics:

	r1	r2	r3	r4	Min(r1-r4)	Min(r1-r2)
L1	1.1	1.0	0.9	1.0	1.0	1.0
L2	0.3	0.9	0.3	0.9	0.9	0.9

Values in these tables equal to 1.0 indicate stability of the results, and values less then 1.0 indicate *improvement* of the results. *Both* cases are **contrary** to McKay *et al*'s thesis.

It should be especially noted that the statistic Min(r1-r4) for L2 also improves. McKay *et al* worked hard to "cook" a set of 135 variations in which this statistic yielded only 4 improvements. But in our article [1] we have already shown that when we check variations that McKay *et al* "forgot" – the results are totally different. Here too, even though we checked the one and only new variation mentioned by McKay in [2], we immediately found an improvement!

This is apparently the reason why McKay *et al* included the variation of example 1 (mentioned above in Chap. I, Sec. A), in their "Study of Variations" but not that of example 2.

C. The statistic Min(P1-P2) bothers McKay because it proves that there was **no** optimization in our original work. McKay tries to escape this implication with baseless arguments. In another reply [7] we already showed that all his contentions are nonsense.

Appendix

[Written on 7 Nissan (20 March 2002) in response to the alterations McKay made in his Appendix].

We recently noticed that McKay has altered his Appendix following our criticism above, where we proved that his Appendix is primarily designed to mislead the reader. His new version has the same goal, albeit this time he tries to be more sophisticated.

A. At the beginning of his original Appendix McKay announced:

> "Here are two examples where the success measure presented by WRR with their first list of rabbis differs mathematically from the success measure presented by WRR with their second list of rabbis. The two scans below are from the 1986 preprint in which WRR presented their first list of rabbis."

But, since we proved (Chap. I, Sec. A) that this claim is empty, McKay replaced it with the following statement:

> "Here are two examples where the success measure presented by WRR in their earliest preprints differ mathematically from the success measure published by them in Statistical Science. The 1986 preprint presented the first list of rabbis, and 1987 preprint presented the second list. Neither preprint mentioned the permutation test that appeared in Statistical Science."

1. Note that the original claim was supposed to create the (mistaken) impression that there is a difference between what McKay calls M1 and what he calls M2. But his new version doesn't mention this at all, and only speaks of the differences in (what he calls) M3!

In Chap. I (Sec. A), we totally destroyed the first version. Therefore, McKay was forced to replace it with this new claim concerning M3, knowing that we already agreed that in some cases there is a difference between M3 and the previous program (see Chap I, Sec. B).

2. Note that McKay adds at the end of the new paragraph:

> "Neither preprint mentioned the permutation test that appeared in Statistical Science."

This statement is correct: It is well known that the permutation test was proposed a long time *after* the preprints were publicized, therefore, it could not have been mentioned in them. Since the permutation test is irrelevant here, and even McKay did not mention it in connection with the "two examples," we must conclude that McKay added this superfluous statement only to create the false impression that something was "not right" with the preprints.

B. The two examples:

1. Concerning what McKay says about the perturbations.

- McKay **admits** that the **same method** was always used (in what he calls M1, M2 and M3) to create the group of perturbations:
  "In this case there is evidence that the last 3 gaps were used in all three computations…"
- But he creates a warped argument to "explain" the relevant description in the second preprint in a twisted way.

Note that McKay started with his "Study of Variations " [6] where he pretended to prove that WRR "cooked" their data. Following our refutation [1] of his "proofs", he was forced to make a "reconsideration of his hypothesis" and was left with speculations [2]. After we disproved even these speculations in this article, he chose to make pointless "inferences" about one preprint or another, even though he admits that *this makes no difference to how the experiment was executed.*

2. Concerning the method used to calculate the maximum skip.

Let us summarize his objections:

- The definitions in the first preprint and in the article in *Statistical Science* differ from one another. But concerning the **actual** method used, he cannot determine whether any changes were made or not:
  "We have been unable to determine which variation was used in the three computations".

- His main claim: The alteration was already made in the second preprint, but it is described unclearly in order to disguise this fact.

Once again, all his "proofs" are based on imaginary "inferences" drawn from our wording, and he admits of having no other evidence.

References

1. Witztum, D., Beremez, Y. (2000). MBBK's Study of Variations. Available at: http://www.torahcode.co.il/english/variate.html.
2. McKay, B. D. (2001). The Analysis of Variations – a reply to Doron Witztum. http://cs.anu.edu.au/ ~ bdm/dilugim.
3. 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.
4. Witztum, D., Rips, E. and Rosenberg, Y. (1986). Equidistant Letter Sequences in the Book of Genesis. Preprint.
5. Witztum, D., Rips, E. and Rosenberg, Y. (1988). Equidistant Letter Sequences in the Book of Genesis. Preprint.
6. 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.
7. Witztum, D. (2001). Smoke Screen: Concerning McKay's Response to our Article "Concerning the Choices of Dates for WRR's Rabbis Samples. Part A: Direct Optimization." Available at: http://www.torahcode.co.il/english/pdf_files/dat2rese.pdf.