New draft, Elul 25, 5758 (Sept. 16, '98).
First draft Iyyar 14, 5758 (May 10, '98).

The "Famous Rabbis" Sample: A New Measurement

By Doron Witztum and Yosef Beremez [note 1]

Introduction:

In the experiment described in [1] significance was measured using a randomization test. This test was first developed for use on the second sample of famous rabbinical personalities (see [1] for details). The purpose of the test was to determine whether the Overall Measures of Proximity for the sample – P_1 and P_2 – are "unusually low." The test compared the values of P_i of the original sample with the values for 999,999 permutated samples, compiled by randomly associating names of personalities in the original sample with dates pertaining to their colleagues.

During the past year this test came under criticism from Dr. B. D. McKay [2]. Dr. McKay claims that the test incorporates a methodological error. We will discuss his assertion and show empirically – using a ***different*** randomization – that the high level of significance received in the original test was not a product of methodological error.

I. Dr. McKay's claim:

Dr. McKay criticized the significance test described in [1], claiming that the test incorporates a methodological error [note 2]. Let us examine his criticism: The sample under study is a set of "name-date" pairs. Suppose that the ELSs of a certain appellation have an "advantage" over its occurrences in PLSs (perturbed letter sequences, where the distance between the letters is unequal). This advantage, for example, may take place if the ELSs occur more frequently. We call this effect "charisma". When we make our calculations for the convergences between the ELSs of this appellation and more than one date (or form of the date), we end up taking advantage of this effect more than once. This would be a methodological error.

Actually, problems of this sort have been addressed already by our randomized pairing test: Suppose that the success of the convergences of a particular appellation was due entirely to its "charisma". If this were the case, this charismatic appellation should succeed equally well with other dates. The results of the permuted sample, in which random pairings replace the correct ones should be succeed to about the same degree. Thus the randomization test should serve to cancel the effects of the charisma of any particular appellation. Dr. McKay, however, claims that residual effects can still have a significant effect on the results.

II. The new measurement:

In our estimation the residual effect mentioned by Dr. McKay is marginal, and only has a negligible effect on the results. To demonstrate this, we subjected the second sample to a different randomization test[note 3]. We reasoned as follows:

1. If a word *w* appears in the sample of word pairs more than once, we can negate any possible advantage it may have. When we calculate *c(w,w')*, we can consider the "first" word *w* only as ELSs, while the "second" word *w'* is taken as ELSs and as PLSs. In other words, the ELSs of *w*' compete with the PLSs of it over

the more successful proximities to the ELSs of *w*. Thus, any charisma that *w* might have will be just as exploited by all the competitors.

2. This strategy solves the problem for the “first” word, but not for the “second”. Therefore, we must arrange that every expression occurring as a “second word” be used no more than once. The sample under investigation consists of word pairs in which one word is the appellation of a rabbi and the second is a date. Usually there is more than one appellation for each personality. If we take the appellation as the “first” word, then we will have the same date as the “second” word several times.

The date of birth or death was used in 3 different forms: א׳ תשרי, בא׳ תשרי, א׳ בתשרי. Therefore, each appellation will, as a rule, take part in 3 pairings, that is, in association with each form of the date. Thus if we take the date as the “first” word, we will have to take each appellation as the “second” word several times.

The solution: Let us divide the sample into three sets: Set 1 – in which the dates are of the form א׳ תשרי; Set 2 – in which the dates are of the form בא׳ תשרי; and Set 3 – in which the dates are of the form א׳ בתשרי.

Let us look, for example, at Set 1: The first personality on our list of rabbis has several appellations: רבי אברהם, הראב״י, הראב״ד, הרב אב״ד, and האשכול. He passed away on the 20th of Cheshvan (כ׳ חשון). We will calculate the convergences of :

רבי אברהם --- כ׳ חשון,
הראב״י --- כ׳ חשון,
הראב״ד --- כ׳ חשון,
הרב אב״ד --- כ׳ חשון,
האשכול --- כ׳ חשון.

In all our calculation we will take the date as the “first” word, and we will take it only as ELSs. However, the ELSs of each appellation (the "second" word) will compete with its PLSs over the more successful proximities to the ELSs of the date. We will follow the same procedure for all the dates and appellations in the sample. In each set every appellation appears only once, with the exception of appellations of the form “Rabbi So-and-So”, which sometimes apply to more than one personality (for example, several personalities were known as “Rabbi Avraham”). To avoid this problem one could, for example, take only that “Rabbi So-and-So” whose date is the first in the sample which appears as an ELS.

3. In this manner we receive a set of results *c(w,w’)*, for which we can then calculate values of P_i. Now we would like to know whether these values are “unusually low.”

4. To this end we will perform the New Randomization: We type the date כ׳ חשון into the computer. When it has registered, we proceed to enter the appellations. But this time instead of processing the name as we typed it, the computer first scrambles the letters of the name, for example, using רהבא״י instead of הראב״י , and only afterwards pairs it with a date. In other words, the letters comprising the expression are subjected to a random permutation. We continue with this procedure for all the pairs in the first set. In each case the “second” word (the appellation) is scrambled by a random permutation. Thus we receive for this perturbed sample a new set of results *c(w,w’)*, for which we will calculate the value of P'_i.

The number of perturbed samples one can construct in this manner is enormous. Let us label it N (one of these samples is the original set 1). Theoretically one could calculate P'_i for all the samples of this sort. We would then have N values for P'_i. We could then arrange these values in order of magnitude. If the phenomenon we are measuring is random, the value P_i (the Overall Measure of Proximity for Set 1) has an equal chance of occupying any of the N positions on the list of values of P'_i.

This is our null hypothesis. It should be noted that this null hypothesis, and the derived significance test, do not make use of any of the considerations which guided us in defining the Corrected Distance and the Overall Measures of Proximity, according to which they had statistical meaning. Therefore, this significance test can be regarded as a “black box” test.

As has been mentioned previously, the number N is enormous. For this reason we were unable to calculate all the values of P'_i for all N samples. In order to determine statistical significance, we will allow the computer to repeat the procedure of compiling perturbed samples M times, where M is some large figure. We will calculate P'_i for each of these samples. Including P_i, we will have M+1 values, which we can then arrange according to the usual order of real numbers. We will define the “rank” of P_i among the M+1 values as the number of P'_i whose magnitude is no greater than that of P_i (if some of the values for P'_i are exactly equal to P_i, we will consider half of them to "exceed" P_i). Next we will define r_i as the rank of P_i divided by M+1. r_i expresses the probability of P_i achieving such a low ranking.

III The results:

We ran the above test using M = 999,999 permuted samples. We recorded the ranking out of 1,000,000 for the values of P_i of each of the sets defined in the previous section:

Table 1

	The Rank of P_1	The Rank of P_2
Set 1	71	2
Set 2	18,777	12,928
Set 3	228,408	5,993

The first set was the most successful, particularly P_2. Therefore we ran an additional test for P_2 of this set using M=999,999,999 permuted samples. Its ranking was 313 out of 1,000,000,000. We calculated r_i and min r_i for each set. The level of significance of each group is 2min r_i.

Table 2

	min r_i	Significance
Set 1	0.000000313	0.000000626
Set 2	0.0129	0.0258
Set 3	0.00599	0.012

IV Conclusion:

Using a completely **different** randomization from that used in [1], we again received an extremely high level of significance. It should now be perfectly clear that the potential defects Dr. McKay noted in the method of randomization had at most a negligible effect. If they had any affect at all on the results of [1], it was a detrimental effect not a positive one.

Appendix

Here are some technical points concerning the measurement above:

1. We took one of the set's pairs and carried out 100 different permutations of the appellation. In the event that the number of possible different permutations n was less than 100, we performed n permutations. The permutations were conducted in a standardized manner using a program designed by Yaakov Rosenberg.

For example: the first pair is "כ׳ חשון - רבי אברהם". We shall present here some of the pairs which are formed by the permutations (by order, from left to right) and also the original pair:

כ׳ חשון רביאברהמ	כ׳ חשון בברמהאי	כ׳ חשון בברמאהי	כ׳ חשון בברמאיה	• • •	כ׳ חשון בבמרראיה	כ׳ חשון בבמרריאה

2. We calculated the values of *c(w,w')* for the convergences of all 100 (or n) permutated appellations with the date taken only as ELS's, as described in sec. 2, par. 4. For example, with regards to the example above, we obtain a row of cells. In each cell there is a *c*-value of the specific pair. An empty cell means that the permutation of the appellation did not appear as ELS:

-	-	49/65	25/71	•••	67/74	62/72

3. If an appellation of one of the personalities is a part of another appellation of his, we took care that this relation will be kept in their permutations as well.

For example: the appellation "מהרח״ש" is included in the appellation "המהרח״ש". The permutations of "המהרח״ש" were taken as the permutations of "מהרח״ש" with a ״ה״ as a prefix. Here are some of the pairs which are formed by the permutations of "מהרח״ש" (by order, from left to right) and also the original pair:

י״ג ניסן מהרחש	י״ג ניסן חשרהמ	י״ג ניסן חשמהר	י״ג ניסן חשהרמ	• • •	י״ג ניסן המרחש	י״ג ניסן המרשח

Their *c*-values are:

37/125	8/125	80/125	46/125	•••	80/125	71/125

In parallel, the permutations for ״המהרח״ש״ give the following:

י״ג ניסן המהרחש	י״ג ניסן החשרהמ	י״ג ניסן החשמהר	י״ג ניסן החשהרמ	• • •	י״ג ניסן ההמרחש	י״ג ניסן ההמרשח

And their *c*-values are:

35/125	120/125	91/125	115/125	• • •	23/125	117/125

We slotted these numbers into one row of cells: in each cell there are two *c*-values: the one for the permutation of ″מהרח״ש" and one for the parallel permutation of "המהרח״ש":

37/125 35/125	8/125 120/125	80/125 91/125	46/125 115/125	•••	80/125 23/125	71/125 117/125

4. Stages 1,2 and 3 were performed with regards to all the pairs in the set. We thus obtained rows of cells, each containing 101 (or n+1) cells. In each cell which is not empty, there are one, two or more values of $c(w,w')$.

5. We then chose by lottery one of the cells in the first row, one of the cells in the second row, and so on. We obtained a set of values of $c(w,w')$ and we calculated the values of P_i for them.

6. We repeated this procedure 999,999 times, using an algorithm for randomization similar to that described in [1]. The program used was also prepared by Yaakov Rosenberg. We used a seed of 10.

7. For Set 1 we ran the lottery 999,999,999 times using the same program and the same seed.

Bibliography:

1. D. Witztum, E. Rips & Y. Rosenberg, "Equidistant Letter Sequences in the Book of Genesis", *Stat. Science*, Vol. 9 (1994), No. 3, pp. 429-438. Available at http://www.torahcode.co.il/pdf_files/pub/wrr.pdf.
2. B. D. McKay, "Equidistant Letter Sequences in Genesis – A Report" (Draft), Apr. 3, 1997.
3. D. Witztum, E. Rips & Y. Rosenberg, "A Hidden Code in the Book of Genesis- the Statistical Significance of the Phenomenon", ("צפן חבוי בספר בראשית") preprint 1996 (Hebrew).

Notes:

Note 1: Literary name.

Note 2: Dr. McKay [2] has also raised other claims against the randomization test: "forget Genesis for a moment and look just at the list of names and dates. They are very varied. The personalities have from 1 to 11 appellations, and from 0 to 6 dates. As well as that, some appellations and dates are short and some are long. Some have letters which are uncommon in Hebrew and some have only common letters. These variations mean that there are many differences between the permutations even before Genesis is considered. A simple example is that the number of name-date pairs varies by more than 100 between different permutations. Why, then, are we justified in assuming they will have comparable a-priori distributions of P_1-P_4 statistics?"
In fact, the new measurement answers these claims as well.

Note 3: This test is essentially the same randomization test that we proposed and implemented in our work on “Headline” samples (see [3]).