## 6. APPELLATIONS FOR *WAR AND PEACE*

An Internet publication by two of the present authors (Bar-Natan and McKay, 1998), presented a new list of appellations for the 32 rabbis of WRR's second list. The appellations are not greatly different from WRR's: 83 were kept, 20 were deleted and 29 additional appellations were added. Many of the changes were simply replacements of one valid spelling by another. The punch line is that the new set of appellations produces a "significance level" of one in a million when tested in the initial 78,064 letters (the length of Genesis) of a Hebrew translation of Tolstoy's *War and Peace*, and produces an uninteresting result in Genesis. Exactly the same text of *War and Peace* is used for control tests in WRR94.

All of our changes were justified either by merely being correct, or by virtue of being no more doubtful than some analogous choice made in WRR's list. For example, whereas WRR used one common Hebrew spelling of the name "Horowitz," we used a different common spelling. When they omitted one common appellation, we inserted it and deleted another. And so on. Our list of appellations does not aspire to be perfect, merely to be of quality commensurate with that of WRR's list. As verified by Menachem Cohen, there is "no essential difference" between WRR's list and ours (Cohen, 1997a). (Amusingly, one knowledgable rabbi who inspected both lists pronounced them "equally appalling.")

This demonstration demolishes the common perception and oft-repeated claim that the freedom of movement left by the rules established for WRR's first list was insufficient by itself to explain an astounding result for the second list.

The appellation list of Bar-Natan and McKay (1998) has been the subject of concerted attack (Witztum, 1998a). The essence of his thesis is that WRR's lists were governed by rules and that the changes made in the second list to tune it to *War and Peace* violate these rules. However, most of these "rules" were only laid out nine to ten years after WRR's two lists were composed, in a lengthy letter written by Havlin (1996) in response to some questions we raised, and had never been publicly mentioned before. While the letter offers many explanations and examples of Havlin's considerations when selecting among possible appellations, they are far from being rules and are fraught with inconsistency. Moreover, when rules for a list are laid out a decade after the lists, it is not clear whether the rules dictated the list selections, or just rationalize them. Besides, as Bar-Natan and McKay amply demonstrate (1999), these "rules" were inconsistently obeyed by WRR.

Most of Witztum's criticisms are inaccurate or mutually inconsistent, as the following two examples illustrate.

1. Witztum argues against our inclusion of some appellations on the grounds that they are unusual, yet defends the use in WRR94 of a signature appearing in only one edition of one book and, it seems, never used as an appellation.
2. Similarly, Witztum defends an appellation used in WRR94 even though it was rejected by its own bearer, on the grounds that it is nonetheless widely used, but criticizes our use of another widely used appellation on the grounds that the bearer's son once mentioned a numerical coincidence related to a different spelling.

These are but two of many examples. Clearly, the issue of the comparative quality of the two lists, which involve historical and linguistic considerations inappropriate to this journal, cannot be broached further here. But Cohen's cited remarks, as well as work to be discussed in Section 10, support our claim to have produced a list no less rule-bound or error-free than WRR's.

Prompted by Witztum's criticisms, we adjusted our appellation list for *War and Peace* to that presented in Table 2. Compared to our original list, it is more historically accurate, performs better, and is closer to WRR's list. Note that we have removed two rabbis who have no dates in WRR's list and one rabbi whose right to inclusion was marginal. We also added one rabbi whom WRR incorrectly excluded and imported the birth date of Rabbi Ricchi in the same way that they imported the birth date of the Besht for their first list. As in WRR94, our appellations are restricted to five to eight letters. Detailed justifications, including responses to Witztum's critique, can be found in our updated paper (Bar-Natan and McKay, 1999), and an associated paper (Anonymous, 1999).

Several more examples of "experiments" performing well in *War and Peace* are mentioned in Section 9.

## 7. THE STUDY OF VARIATIONS

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. Since WRR are claiming what can only be described as statistical proof of a miracle, the presence of so much "wiggle room" in the design, together with our failure to obtain any support for their claims from our

TABLE 2
*Appellations for War and Peace*

| Personality | Appellations |
|---|---|
| Rabbi Avraham Av-Beit-Din | רבי אברהם, הראב״י, הרב אב״ד, הראב״ד, האשכול |
| Rabbi Avraham Yitzhaki | רבי אברהם, יצחקי, זרע אברהם |
| Rabbi Avraham Ha-Malakh | רבי אברהם |
| Rabbi Aaron of Karlin | רבי אהרן |
| Rabbi Eliezer Ashkenazi | מעשי השם, מעשי י/ה/ו/ה, מעשי ה׳, בעל מעשי ה׳ |
| Rabbi David Oppenheim | רבי דוד, אופנהים |
| Rabbi David Nieto | רבי דוד, דוד נישו |
| Rabbi Chaim Abulafia | רבי חיים, המהרח״א, מהרח״א |
| Rabbi Chaim Benbenest | רבי חיים, בנבנשתי, הרב חבי״ב, הרב החבי״ב, רב חבי״ב |
| Rabbi Chaim Capusi | רבי חיים, כאפוסי |
| Rabbi Chaim Shabtai | רבי חיים, חיים שבתי, מהרח״ש, המהרח״ש |
| Rabbi Yair Chaim Bacharach | חות יאיר |
| Rabbi Yehudah Chasid | רבי יהודה, יהודה סג״ל, הר״י חסיד |
| Rabbi Yehudah Ayash | רבי יהודה, מהר״י עיאש, עאיאש |
| Rabbi Yehosef Ha-Nagid | רבי יהוסף |
| Rabbi Yehoshua of Cracow | רבי יהושע, מגני שלמה |
| The Maharit | רבי יוסף, מטרני, יוסף טרני, טראני, מטראני, מהרימ״ט, המהרימ״ט<br>מהרי״ט, המהרי״ט, הר״י טרני, הר״י טראני, ר״י טרני, ר״י טראני |
| Rabbi Yaacov Beirav | רבי יעקב, יעקב בירב, מהר״י בירב, הריב״ד |
| Rabbi Israel Yaacov Chagiz | בעל הלק״ט, מהר״י חגיז, ר״י חגיז |
| The Maharil | רבי יעקב, מולין, יעקב סג״ל, יעקב הלוי, מהר״י סג״ל, מהר״י הלוי, מהרי״ל, המהרי״ל |
| The Yaabez | היעב״ץ, הריעב״ץ, עמדין, הר״י עמדין, ר״י עמדין |
| Rabbi Yitzhak Ha-Levi Horowitz | רבי יצחק, הורוביץ, יצחק הלוי |
| Rabbi Menachem Mendel Krochmal | רבי מנחם, קרוכמאל, רבי מענדל, צמח צדק |
| Rabbi Moshe Zacut | רבי משה, משה זכות, מהר״ם זכות, מהרמ״ז, המהרמ״ז, המזל״ן, קול הרמ״ז |
| Rabbi Moshe Margalith | רבי משה, מרגלית, פני משה, מרגליות |
| Rabbi Azariah Figo | רבי עזריה |
| Rabbi Immanuel Chai Ricchi | הון עשיר, העשי״ד, אוהב ור״ע |
| Rabbi Shalom Sharabi | רבי שלום, שרעבי |
| Rabbi Shlomo of Chelm | רבי שלמה, שלמה חלמא, חעלמא |
| Rabbi Meir Eisenstat | רבי מאיר, איזנשטט, איזנשטאט, מהר״ם א״ש |

own experiments (detailed in Section 10), should be sufficient reason in itself to disregard WRR's findings. However, one can do more: there is significant circumstantial evidence that WRR's data is indeed selectively biased toward a positive result. We will present this evidence without speculating here about the nature of the process which led to this biasing. Since we have to call this unknown process something, we will call it *tuning*.

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 (by which we mean that there was no clear reason under WRR's research hypothesis that they should be made in the particular way they chose to) and see how often these decisions turned out to be favorable to WRR.

### Direct Versus Indirect Tuning

We hasten to add that we are not claiming that WRR tested all our variations and thereby tuned their experiment. This naturally raises the question of what insight we could possibly gain by testing the effect of variations which WRR did not actually try. There are two answers. First, if these variations turn out to be overwhelmingly unfavorable to WRR, in the sense that they make WRR's result weaker, the robustness of WRR's conclusions is put into question whether or not we are able to discover the mechanism by which this imbalance arose. Second, and more interestingly, 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. Fortunately for our analysis, we do not need to distinguish which possibility holds in each case. (However, we note that for the first list practically all aspects of the experiment were available for tuning, while for the second list many features had been fixed by the first list. The primary possibility for tuning of the second list was in appellation selection, but some aspects of the test method were free too.)

### The Space of Possible Variations

Our approach will be to consider only minimal changes to the experiment. An inexact but useful model is to consider the space of variations to be a direct product $X = X_1 \times \cdots \times X_n$, where each $X_i$ is the set of available choices for one parameter of the experiment. The model supposes that the choices could be applied in arbitrary combination, which will be close to the truth in our case. Call two elements of $X$ *neighbors* if they differ in only one coordinate. Instead of trying to explore the whole (enormous) direct product $X$, we will consider only neighbors of WRR's experiment in each of the coordinate directions.

To see the value of this approach, we give a tentative analysis in the case where each parameter can only take two values. For each variation $x = (x_1, \ldots, x_n) \in X$, define $f(x)$ to be a measure of the result (with a smaller value representing a stronger result). For example, $f(x)$ might be the permutation rank of $P_4$. A natural measure of optimality of $x$ within $X$ is the number $d(x)$ of neighbors $y$ of $x$ for which $f(y) > f(x)$. Since the parameters of the experiment have complicated interactions, it is difficult to say exactly how the values $d(x)$ are distributed across $X$. 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. Such functions $f$ are called *completely unimodal* (Ziegler, 1995, page 283). In this case, it can be shown that, for the uniform distribution on $X$, $d(x)$ has the binomial distribution $\mathrm{Binom}(n, 1/2)$ and is thus highly concentrated near $n/2$ for large $n$ (Williamson Hoke, 1988).

Of course, this analogy only serves as a rough guide. In reality, some of the variations involve parameters that can take multiple values or even arbitrary integer values. A few pairs of parameter values are incompatible. And so on. In addition, one can construct arguments (of mixed quality) that some of the variations are not truly "arbitrary." For these reasons, and because we cannot quantify the extent to which WRR's success measures are completely unimodal, we do not 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.

### 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. 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). Most of our variations will indeed be of that form.

### Computer Programs

A technical problem that gave us some difficulty is that WRR have been unable to provide us with their original computer programs. Neither the two programs distributed by WRR (Rosenberg, undated), nor our own independent implementations of the algorithm as described in WRR's papers (1986, 1987, 1994), consistently produce the exact distances listed in those preprints or the histograms that appear there and in WRR94. Consequently, we have taken as our baseline a program identical to the earliest program available from WRR, including its half-dozen or so programming errors. As evidence of the relevance of this program, we note that it produces the exact histograms given in WRR94 for the randomized text $R$, for both lists of rabbis. (The histograms for Genesis that appear in WRR94 are, according to Witztum, the results of a program, presumably lost, that preceded the one used for the permutation tests in WRR94.)

### 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? As noted in Section 3, WRR's success measures varied over time and, until WRR94, consisted of more than one quantity. We will restrict ourselves to four success measures, chosen for their likely sensitivity to direct and indirect tuning, from the small number that WRR used in their publications.

In the case of the first list, the only overall measures of success used by WRR were $P_2$ and their $P_1$-precursor (see Section 3). The relative behavior of $P_1$ 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, $P_2$ depends continuously on all of the $c(w, w')$ values, so it should make a more sensitive indicator of tuning. Thus, we will use $P_2$ for the first list.

For the second list, $P_3$ is ruled out for the same lack of sensitivity as $P_1$, leaving us to choose between $P_2$ and $P_4$. These two measures differ only in whether appellations of the form “Rabbi X” are included ($P_2$) or not ($P_4$). 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 $P_4$ to be a more sensitive indicator of tuning than $P_2$. Thus, we will use $P_4$. Our choice notwithstanding, we feel that $P_4$ imperfectly captures WRR's probable intentions. For their experiment on the second list to have been as successful as first reported (WRR, 1986), WRR needed more than just a small value for $P_2$ or $P_4$. They also needed the distances for a cyclic shift of the dates to show a flat histogram and yield a *large* value of $P_2$ or $P_4$.

In addition to $P_2$ for the first list and $P_4$ for the second, we will show the effect of experiment variations on the least of the permutation ranks of $P_{1\text{–}4}$. This is not only the sole success measure presented in WRR94, but there are other good reasons. The permutation rank of $P_4$, for example, is a version of $P_4$ 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 $P_4$ should often be a more reliable indicator of tuning than $P_4$ itself. The permutation rank also to some extent measures $P_{1\text{–}4}$ for both the identity permutation and one or more cyclic shifts, so it might tend to capture tuning toward the objectives mentioned in the previous paragraph. (Recall from Section 3 that WRR had been asked to investigate a “randomly chosen” cyclic shift.)

In summary, we will restrict our reporting to four quantities: the value of $P_2$ for the first list, the value of $P_4$ for the second list, and the least permutation rank of $P_{1\text{–}4}$ for both lists. In the great majority of cases, the least rank will occur for $P_2$ in the first list and $P_4$ in the second.

### The Results

Values for each of these four measures of success will be given as ratios relative to WRR's values. A value of 1.0 means “less than 5% change.” Values greater than 1 mean that our variation gave a less significant result than WRR's original method gave and values less than 1 mean that our variation gave a more significant result. Since we used the same set of 200 million random permutations in each case, the ratios should be accurate to within 10%. To save space with large numbers, we use scientific notation; for example $3e7$ means $3 \times 10^7$. The score given to each variation has the form $[p_1, r_1; p_2, r_2]$, where

$p_1 =$ The value of $P_2$ for the first list, divided by $1.76 \times 10^{-9}$;

$r_1 =$ The least permutation rank for the first list, divided by $4.0 \times 10^{-5}$;

$p_2 =$ The value of $P_4$ for the second list, divided by $7.9 \times 10^{-9}$;

$r_2 =$ The least permutation rank for the second list, divided by $6.8 \times 10^{-7}$.

These four normalization constants are such that the score for the original metric of WRR is [**1**, **1**; **1**, **1**]. A bold "**1**" indicates that the variation does not apply to this case so there is necessarily no effect.

Two general types of variation were tried. The first type involves the many choices that exist regarding the dates and the forms in which they can be written. A much larger class of variations concerns the metric used by WRR, especially the complicated definition of the function $c(w, w')$. In both cases the details are quite technical, so we have presented them in Appendix B and Appendix C, respectively. Our selection of variations was in all cases as objective as we could manage; we did not select variations according to how they behaved. 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.

### 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. At the very least, this trend shows WRR's result to be not robust against variations. Moreover, as explained at the beginning of this section, we believe that these observations are strong evidence for tuning, but will not attempt a quantitative evaluation.

## 8. TRACES OF NAIVE STATISTICAL EXPECTATIONS

There are some cases in the history of science where the integrity of an empirical result was challenged on the grounds that it was "too good to be true" (Dorfmann, 1978; Fisher, 1965, for example); that is, that the researchers' expectations were fulfilled to an extent which is statistically improbable. Some examples of such improbabilities in the work of WRR and Gans (Gans, 1995, described in Section 9) were examined by three of the present authors (Kalai, McKay and Bar-Hillel, 1998). Here we will summarize this work briefly. It is worthy of note that these observations are surprising even if we adopt WRR's hypothesis that the codes are real.

Our interest was roused when we noticed that the $P_2$ value (not the permutation rank, which did not yet exist) first given by WRR for the second list of rabbis (WRR, 1987), $1.15 \times 10^{-9}$, was quite close to that of the first, $1.29 \times 10^{-9}$. To see whether this was as statistically surprising as it seemed, we conducted a Monte Carlo simulation of the sampling distribution of the ratio of two such $P_2$ values. This we did by randomly partitioning the total of 66 rabbis from the two lists into sets of size 34 and 32—corresponding to the size of WRR's two lists—and computing the ratio of the larger to the smaller $P_2$ value for each partition. Although such a random partition is likely to yield two lists that have more variance within and less variance between than in the original partition (in which the first list consisted of rabbis generally more famous than those in the second list), our simulation showed that a ratio as small as 1.12 occurred in less than one partition in a hundred. (The median ratio was about 700.)

Even under WRR's research hypothesis, which predicts that both lists will perform very well, there is no reason that they should perform equally well. This ratio is not surprising, though, if it is the result of an iterative tuning process on the second list that aims for a "significance level" (which $P_2$ was believed to be at that time) which matches that of the first list. Nevertheless, our observation was a posteriori so we are careful not to conclude too much from it.

An opportunity to further test our hypothesis was provided by another experiment that claimed to find "codes" associated with the same two lists of famous rabbis. The experiment of Gans (1995) used names of cities instead of dates, but only reported the results for both lists combined. Using Gans' own success measure (the permutation rank of $P_4$), but computed using WRR's method, we ran a Monte Carlo simulation as before. The two lists gave a ratio of $P_4$ permutation ranks as close or closer than the original partition's in less than 0.002 of all random 34-32 partitions of the 66 rabbis.

Previous research by psychologists (Tversky and Kahneman, 1971; Kahneman and Tversky, 1972) has shown that when scientists replicate an experiment, they expect the replication to resemble the original more closely than is statistically warranted, and when scientists hypothesize a certain theoretical distribution (e.g., normal, or uniform), they expect their observed data to be distributed closer to the theoretical expectation than is statistically war-