ב"ה טיוטה כ"ז תשרי התשס"א

על "מחקר הווריאציות" של MBBK

דורון ויצטום ויוסף ברמז

מבוא

הניסוי המפורסם של WRR בעניין הצופן בספר בראשית, אשר פורסם ב- *Statistical Science* [1], משמש נושא למאמר הביקורתי של MBBK (מקי, בר-נתן, בר-הלל, קלעי): "Solving the Bible Code Puzzle" [2], שהתפרסם באותו כתב-עת. מבקרים אלה מנסים להראות כי תוצאת הניסוי של WRR אינה תקפה. במבוא למאמרם כותבים MBBK:

> "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)

בנוגע לשאלה הראשונה MBBK טוענים כי

> "The first question is answered affirmatively in Section 6…"

אבל אין זה כך. לדעתנו, טענה זו כלל אינה נכונה והעבודה אותה הם מציגים בפרק 6 ממאמרם אינה אלא מטעה. מכיוון שנושא זה אינו בתחום הסטטיסטי, ואילו המאמר הנוכחי נועד לעסוק בהיבטים הסטטיסטיים של הדיון, אנו מפנים את הקורא למאמר [3] המטפל במיוחד בטענתם זו.

בנוגע לשאלה השניה כותבים שם MBBK:

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

כוונתם כאן ל"מחקר הווריאציות" שלהם, המהווה חלק מרכזי במאמרם, ונועד לברר האם ישנן ראיות עקיפות לכך ש – WRR "תפרו" את הרשימה השניה של שמות וכינויים כך שתתקבל הצלחה בניסוי שלהם בספר בראשית. מאמרנו זה יתרכז בביקורת על "מחקר הווריאציות" של MBBK.

כך מתארים MBBK את גישתם הבסיסית:

> "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 favourable to WRR." (Pg. 158)

MBBK יודעים יפה, שהניסוי של WRR נעשה תחת אילוצים כך שאי אפשר היה לשנות את תנאי הניסוי והפרמטרים שלו, מכפי שעשו WRR בניסוי הקודם שלהם. לכן, מתבססים MBBK על ההשערה הבאה:

> "…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)

כדי ש"מחקר הווריאציות" של MBBK יהיה בעל משמעות מדעית נחוץ:
(א) לבסס את השערת המחקר.
(ב) להשתמש באוסף לא מוטה של ווריאציות בלתי תלויות.

כשל ב**(א)** מבטל את הערך של העבודה.
כשל ב**(ב)** מבטל לא רק את הערך של העבודה, אלא אף מטיל צל כבד על היושר וההגינות של מבצעיה.

אפילו אם נניח לצורך הדיון שעצם המחקר של MBBK הוא בעל משמעות מדעית, העדרו של אוסף אובייקטיבי וסגור של ווריאציות גורם לכך שמשמעות הניסוי היא אחת משתים:
(1) ראייה לכך שהיה "tuning" בנתונים ששימשו לניסוי של WRR.
(2) ראייה לכך שהיה "tuning" בווריאציות ששימשו ל"מחקר הווריאציות" של MBBK.

במאמרנו זה ננסה להראות כי המסקנה הנובעת מתוצאות "מחקר הווריאציות" של MBBK היא **(2)** ולא **(1)**.

- בפרק **א**, נמנה את הליקויים החמורים בעבודה של MBBK מן הבחינה הלוגית והסטטיסטית. יודגש, שדי בכל ליקוי כשלעצמו לבטל את ערכה של עבודתם.
- בפרק **ב**, נביא דוגמאות לטעויות והטעיות חמורות מן הבחינה המתמטית-סטטיסטית.
- בפרק ג, נראה כי MBBK מציגים במאמרם רק **חלק** מן המדידות שערכו, וכי הדרך בה בחרו להציג את התוצאות מסלפת באופן חמור את התמונה **האמיתית** העולה מן הווריאציות שדווחו על ידם. בפרק זה גם נבהיר מדוע התירוצים הא-פוסטריוריים להצגה החלקית והא-פוסטריורית של התוצאות – אינם תקפים.
- בפרק **ד**, נעמיד את התיזה שלהם בכמה ניסויי ביקורת. למשל, נבחן את התיזה שלהם לגבי רשימה "תפורה" – הרשימה שהם עצמם "תפרו" כדי להצליח ב"מלחמה ושלום". ניסוי זה הוא לפי פרדיקציה שהם עצמם הציבו [4]. פרדיקציה זו טוענת כי התוצאות של רשימתם ב"מלחמה ושלום" צפויות להתקלקל ו/או להשתפר באותה המידה כמו רשימת WRR. אבל:

מתברר, שהפרדיקציה שלהם **נכשלה**, והתוצאות הניסיוניות מפריכות את התיזה שלהם: התוצאות ב"מלחמה ושלום" התקלקלו עקב הווריאציות רק בפחות ממחצית הווריאציות!

בפרק זה נוכיח כי תוצאת "מחקר הווריאציות" היא תולדה של "תפירה" (tuning) של הווריאציות.

- בפרק **ה** נפרוש לפני הקורא את מהלך "האבולוציה של מחקר הווריאציות". "מחקר הווריאציות" עבר תהליך אבולוציוני ממושך בן ארבעה שלבים (לפחות). החוקרים שינו מדי פעם את אוסף הווריאציות, והחליפו באופן **א-פוסטריורי** את צורת ההצגה של תוצאות המדידה, ובכל שלב שעשו זאת, שיפרה צורת ההצגה החדשה (אפילו לגבי אותן ווריאציות) את התוצאה לה ייחלו. בפרק זה נביא ראיות ניסיוניות נוספות ל"תפירה" שנקטו בה, ולכך שאין שום קשר בין התוצאות שהציגו MBBK לבין "אופטימיזציה" של הנתונים בניסוי WRR.

פרק א. על הליקויים הלוגיים והסטטיסטיים בתזה של MBBK

כאן נדון בתזה שהמציאו MBBK לצורך "מחקר הווריאציות" מן ההיבט הלוגי והסטטיסטי. עיקר הביקורת שלנו בתחומים אלה מתרכזת בנקודות הבאות:

1. "מחקר הווריאציות" נועד להבחין בין שתי אפשרויות הבאות:

א. האפשרות עליה מצביעים WRR והיא, שקיים בבראשית צופן בדילוגים שווים.

ב. האפשרות עליה מצביעים MBBK, והיא, שאין הוכחה שיש בבראשית צופן בדילוגים שווים, וכי ההצלחה של WRR נבעה מ"תפירה" של רשימת מלים לניסוי (רשימת השמות וכינויים).

אם קיימת תופעה, יש לה מאפיינים מסוימים, ולכן סביר מאד שדווקא ניסויים הבנויים בהתאם למאפיינים אלה יצליחו. למשל, בניית הפונקציה $c(w,w')$ המודדת את הקירבה במפגשים, נעשתה תוך ניסיון להשתמש במאפיינים שאובחנו במפגשים של זוגות מלים בדוגמאות קודמות. הפונקציה שנבחרה משקפת, למשל, את העובדה שאובחנו f ו- l קטנים בדוגמאות הקודמות (ראה בנספח, פרק א). אין לנו ספק, שפונקציה זאת אינה יחידאית. אולם סביר הדבר, שפונקציות אחרות שיצליחו ישקפו אותם מאפיינים שאובחנו בדוגמאות הקודמות.

מצד שני, לשיטתם שאין תופעה, הרי ההצלחה של WRR הושגה על ידי "תפירה" ביחס לפונקציה זו דווקא. ולכן גם לשיטתם צפויה לרשימה ה"תפורה" הצלחה רק בפונקציה זו, או בפונקציה הדומה לה.

נובע מכך, שעבור וריאציות רבות אם "מחקר הווריאציות" יגלה שאומנם התוצאה התקלקלה, אי אפשר לדעת אם זה משום **א** או משום **ב**, כיוון <u>שבכל מקרה</u> צפויה התוצאה להתקלקל: אם משום שיש תופעה התואמת מאפיינים מסוימים, אם משום שהיתה "תפירה" ביחס לאותם מאפיינים. **זה ליקוי יסודי של "מחקר הווריאציות".**

2. MBBK מנסים לתקן את הליקוי היסודי הזה על ידי... טעות יסודית נוספת. הבה, ונעיין במאמרם:

> **"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, emphasis ours).

בקטע המודגש הם טוענים שבהשערת אפשרות 1א (שיש תופעה),

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

זו **טעות יסודית**. כפי שהסברנו ב- 1, אם קיימת תופעה, יש לה מאפיינים מסוימים. לכן סביר מאד שדווקא ניסויים הבנויים בהתאם למאפיינים אלה יצליחו.

גם המשך דבריהם שם בנויים על אותה טעות:

> "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)."

מדבריהם כאן ברור, שהדרישה כי הווריאציה לא תחרוג מהגדרת התופעה אינה **הכרחית** בעיניהם, אלא נועדה רק להוכיח <u>ביתר בירור</u> את טענתם. אם הם היו מבינים שדרישה זו היא בעצם **תנאי הכרחי**, לא היו נכשלים כאן בשני עניינים חשובים:

(א) הם לא הגדירו במונחים <u>מדויקים</u> מה נכלל בתופעה ומה חורג ממנה. ביטויים מעורפלים כמו

"the concept underlying the list of word pairs and the informal notion of ELS proximity",

רק מאפשרים להם לכתוב בהמשך:

> "Most of our variations will indeed be of that form."

כי בשק הזה אפשר להכניס כמעט כל דבר. כשל זה בולט במיוחד, לאור דברי מקי שכתב בדו"ח שלו [5] (ממנו נלקחו מרבית הווריאציות), את הדברים הבאים:

> "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 **<u>is highly subjective</u>**". (Emphasis ours).

(ב) הם לא הקפידו על כך **שכל** הווריאציות יהיו מסוג זה, ולא רק **רוב** הווריאציות. לאור דבריהם לעיל, אנו מבינים עכשיו מדוע MBBK הרשו לעצמם לכלול ווריאציות **שאינן בתחום התופעה** אף לפי דעתם. (אגב, לעומתם אנו סבורים, **שרוב** הווריאציות שלהם הן דווקא מחוץ לתחום התופעה).

על כל פנים רשמנו לפנינו את **שתי הודאותיהם**:
• כי חלק מן הווריאציות שלהם חורגות מתיאור התופעה על ידי WRR, ולכן הן מן הסוג שצפוי **מראש** דווקא לקלקל את התוצאה.
• כי גם הגדרת שאר הווריאציות כ"דומות לניסוי המקורי" היא "highly subjective".

3. הם טוענים בפרק 4 במאמרם, כי תוצאת WRR השתפרה על ידי פלוקטואציה. לשיטתם, שקיימת פלוקטואציה <u>המשפרת</u> את התוצאה, הרי זה מבטיח <u>מראש</u> שתוצאות הווריאציות יצביעו על קילקול ברוב המקרים.

בפרק **ב** נדגים כיצד השתמשו בכך MBBK לקדם את מטרתם.

4. עיקר טענתם, כי "אופטימיזציה על רשימת המלים (הכינויים) צריכה להראות כמו אופטימיזציה על הפרמטרים של הניסוי" היא בגדר **סברה**. אבל, אנו עוסקים במדע ובמתמטיקה ולכן סברות צריך **להוכיח**. המפליא הוא, ש-MBBK לא ראו צורך לנסות להוכיח הנחה זו. אבל, ראה בפרק **ד** לקמן, כי תוצאות ניסויים שנועדו לבדוק את התיזה של MBBK מטילות ספק כבד בתקפות הנחה זו.

5. MBBK מרבים לטעון כי ביצעו:

- "<u>minor variations</u> on WRR's experiment" (Pg. 152).
- "Our approach will be to consider only <u>minimal changes</u> to the experiment" (Pg. 159).
- "However, since almost all the variations we try amount to only <u>small changes</u> 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).

אך בשום מקום לא הגדירו א-פריורית למה הכוונה בכל אחד מן המונחים:
"minor variations", "minimal changes", "small changes", "natural minor variations to the experiment".
בהעדר קריטריונים כאלה קשה להתייחס ברצינות לאמירות מסוג זה. ואומנם, כפי שנדגים להלן (בפרק **ב**):

א. שינויים רבים שעשו כלל אינם קטנים ובוודאי לא מינימליים.

ב. במקרים שבדקנו שינויים קטנים מאלה של MBBK, אכן קיבלנו תמונה **שונה** לחלוטין.

6. בעיה מרכזית בניסויים מסוג "מחקר הווריאציות" של MBBK היא **התלות** בין הווריאציות: בין אם המדובר בתלות בין הפונקציות שנבחרו לצורך זה, ובין המדובר בנקודות דגימה עבור פרמטר מסוים. למעשה, רובן של הווריאציות המוצגות על ידי MBBK לוקות בחיסרון זה. גם הם מבינים זאת, ולכן אינם טוענים שבאפשרותם לכמת את התוצאות.
אבל למרות **שאי אפשר לכמת את התוצאות** הם משתמשים בהן כדי **להשפיע מבחינה פסיכולוגית**. ראה להלן סעיף 10.
היות ושם המשחק "פסיכולוגיה", ישנו תפקיד מרכזי לצורת הצגת הנתונים. במצב כזה, לכל הטעיה בהצגת הנתונים יש השפעה רבה על הקורא. ראה דוגמא לכך בפרק **ב**, וראה בהרחבה בפרק **ג**.

7. רשימת הווריאציות של MBBK אינה סגורה.

(א) איננו יודעים כמה ווריאציות הם ניסו באמת, וכמה ווריאציות הם זרקו ל — "waste basket" (MBBK הודו שבדקו מאות ווריאציות [6]). אומנם, MBBK קוראים לנו להאמין להם כי

> "Nothing we have chosen to omit tells a story contrary to the story here." (Pg. 159)

אבל, כבר הוכחנו [7] [8], וגם נוכיח להלן במסגרת מאמר זה, כי אי אפשר לתת אמון בהצהרה זו. יתר על כן, אחד ממחברי המאמר כבר הודה בפומבי (ראה להלן פרק **ה** 1(**ג**)(1)) כי אכן נזרקו ווריאציות ל-"waste basket" לאחר שבוצעה לפיהן מדידה, שתוצאותיה שימשו בסיס לטענות נגד WRR – טענות אשר הופרכו לאחר מכן.

(ב) רשימה שאינה סגורה אינה מאובטחת לא רק מהסתרת ווריאציות, אלא גם **מהוספת** ווריאציות. הכוונה לבדיקה א-פוסטריורית של ווריאציות לאור ידע קודם מה מצליח ומה נכשל. בפרק **ב** נראה דוגמאות רבות לכך מרשימת הווריאציות של MBBK. בפרק **ה** נביא ראיה סטטיסטית שהוספה זו היתה מכוונת להצלחת התזה שלהם.

כאן המקום להדגיש כי כבר לאחר השלב הראשון ב"אבולוציה של הווריאציות" כתב המתמטיקאי פרופ' ישראל אומן לפרופ' בר-הלל [9] את הדברים הבאים:

```
"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."

אבל, למרות זאת לא פעלו MBBK לפי דרישה זו שהיא כה אלמנטרית בעבודה מדעית. במקום לעשות זאת, הם פונים אל הקורא בבקשה שיתן בהם אמון:

> "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)

כלומר, הם אומרים: "תאמינו לנו שלא "תפרנו" ווריאציות, ומכח **אמונה** זו אל תאמינו ל-WRR שלא "תפרו" כינויים". לכן, כל "מחקר הווריאציות" מבוסס על אמון ולא על מדע.

8. ב"מחקר הווריאציות" יש עירבוב בין שני נושאי מחקר מדעי השונים זה מזה: ווריאציות ורפליקציות.

(א) **רפליקציה** במהותה היא בדיקת התנהגות התופעה. מניחים הנחות על התופעה – ובודקים אותן. למשל, במקרה שלנו, אפשר להניח שהתופעה מתרחשת לא רק בספר בראשית אלא גם בספר שמות; שהיא מתקיימת גם לגבי צורות תאריך אחרות וכד'. בניסויי רפליקציה משנים את הנתונים הבסיסיים: כגון, שמות חדשים, תאריכים חדשים, ספר חדש או אוכלוסיה חדשה של ELSs.

ניתן לבצע את המחקר של MBBK מתוך <u>הנחות</u> על התופעה. לדוגמא, אפשר להניח שאם התופעה קיימת בספר בראשית היא צריכה להיות קיימת אף בספר שמות. אולם סתירת ההנחה הזאת <u>לא תוכיח</u> ש- WRR זייפו; אלא רק תוכיח שהנחת MBBK על התופעה אינה נכונה. הדיון "מדוע" בספר בראשית ישנה התופעה הזו ולא בספר שמות, אינו בתחום הסטטיסטי אלא בתחום המטאפיסי.

לכן כל ניסוי בו קיימת <u>הנחה</u> על התנהגות התופעה אינו הכלי הנכון לבדיקת טענת האופטימיזציה.

(ב) בפרט, כל "הווריאציות" של MBBK המוצגות בניספח B במאמרם, כלל אינן יכולות להיכלל ב"מחקר הווריאציות". הנחת היסוד המתימטית לגבי ווריאציות היא, שהתופעה אינה תלויה בפרמטר עליו עורכים את הווריאציה. ואז, טוענים MBBK, אנו מצפים ל- 50% שיפורים ול- 50% קילקולים עקב השינויים בפרמטר.

אבל, **אי אפשר** להניח הנחת יסוד זו לגבי "הווריאציות" בניספח B. "ווריאציות" אלו מבוססות על שינויים **בנתונים**. התופעה – שהיא קיומו של צופן בספר בראשית – בוודאי צפויה להיות תלויה בנתונים. טיבו של **צופן** שהוא תלוי חזק בטיב הנתונים: אם נכתוב נתונים לא נכונים, או נשתמש בצורות לשוניות שגויות או כאלו שאינן קיימות בצופן, בוודאי נקבל שינוי דרמתי בתוצאה. לכן, "ווריאציות" אלו אינן אלא רפליקציות, שהדיון בהן הוא בתחום הביבליוגרפיה התורנית ובתחום הלשוני ולא במסגרת "מחקר הווריאציות".

הרפליקציות נמצאות כמעט כולן בנספח B במאמרם, ואנו נטפל בהן במאמר נפרד [10].

9. הבחירה של MBBK בארבע סטטיסטיקות מסויימות כדי להציג את תוצאות הווריאציות (הקטע בשם "What measures should we compare" בעמ' 160), היא בחירה **א-פוסטריורית** של **חלק** מן המדידות. הם בחרו רביעיה אחת, שהיא צרוף אחד מתוך מיליוני צרופים אפשריים (ראה בניספח), על סמך סיפורים א-פוסטריוריים. זה ליקוי עקרוני.

יתר על כן, להלן נראה שהמדובר לא רק בליקוי עקרוני אלא בהטעיה של ממש.

בפרק **ב** נראה כי בחירה א-פוסטריורית זו אפשרה להם להסתיר עובדות חשובות. בפרק **ג**

נציג את שאר המדידות, ואז יתברר לקורא כי התמונה האמיתית מראה את ההפך מטענתם: שאפילו לשיטתם, **לא היתה אופטימיזציה!**

10. "מחקר הווריאציות" **חסר כימות**. הם כותבים:

> "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)

קשה לנו להבין כיצד עבודה החסרה כימות פורסמה בעיתון סטטיסטי: זה ליקוי יסודי מבחינה מדעית-סטטיסטית.

<u>לסיכום:</u>
אם רוצים לערוך מחקר מסוג "מחקר הווריאציות" כדי לגלות תהליך של אופטימיזציה, חייבים קודם כל:

א. להוכיח את תקיפות המודל (ראה סעיף 4 לעיל).
ב. לבנות כלים המבחינים בין תוצאה של אופטימיזציה, לבין תוצאה של צופן או של פלוקטואציה (ראה 1 ו-3 לעיל).
ג. לדאוג לכך, שאוסף הווריאציות יעמוד **בכל אחד** מן התנאים הבאים:
   1. אי חריגה מתחום התופעה. (ראה 1, 2 ו- 5 לעיל).
   2. אי חריגה מתנאי הניסוי המקוריים. (ראה 8).
   3. א-פריוריות וסגירות. (ראה 7 ו- 9 לעיל).
   4. אי תלות. (ראה 6 לעיל).
ד. לספק ניתוח כמותי של התוצאות. (ראה 10 לעיל).

העבודה שהציגו MBBK אינה עונה ולו על אחד מן התנאים הנ"ל.

עד כאן הצבענו על ליקויים יסודיים ב"מחקר הווריאציות" של MBBK הן מבחינת הביסוס הלוגי שלו, והן מבחינת הביסוס הסטטיסטי. לדעתנו, די בחלק מליקויים אלה כדי לזרוק את כל המחקר הזה לאותו "waste basket" אליו הם כבר השליכו את מקצתו. ואומנם, כפי שנראה להלן בפרקים **ד** ו**ה**, התזה של MBBK נכשלת לחלוטין כאשר מעמידים אותה בניסויי ביקורת.

אך לפני שנעשה זאת, נברר את השאלה הבאה: ראינו לעיל כי MBBK הזמינו את הקורא לשפוט את התוצאות באופן אינטואיטיבי. האם דאגו MBBK לספק לקורא תמונה נאמנה כדי לאפשר לו לעשות זאת? האם דאגו MBBK לספק לקורא נתונים אמינים, שנאספו בדגימה נכונה ובלתי מוטה, והמוגשים בהצגה שאינה מסולפת?
את תשובתנו על שאלה זו נציג בפרקים **ב** וג.

פרק ב. טעויות והטעיות ב"מחקר הווריאציות" של MBBK

בפרק זה נדגים שגיאות סטטיסטיות ומתמטיות חמורות שנעשו על ידי MBBK. לשם הבהירות קיבצנו את הדוגמאות לשגיאות בשני מדורים: 1. מדור ההטעיות. 2. מדור הטעויות החמורות.

1. <u>ההטעיות</u>:

(א) בטבלה 5 (עמ' 169 במאמרם) בטור המרכזי בוחנים MBBK תוצאות 33 פונקציות שהן וריאציות של פונקציה *delta* = ה"מרחק" בין שני ELSs. כל הפונקציות הן בחזקה שניה, כמו הפונקציה המקורית של WRR. בטור הימני באותה טבלה הם מציגים תוצאות של 34 וריאציות נוספות: אחת מהן היא השורש הריבועי של הפונקציה המקורית, ושאר 33

הווריאציות נוצרו על ידי הוצאת שורש ריבועי מ- 33 הפונקציות הנ"ל. דהיינו, סך הכל 34 פונקציות נוספות, כולן בחזקה הראשונה.

למתבונן בטור התוצאות הימני מזדקרת לעין עובדה בולטת מאד: **בכל** 68 התוצאות לרשימה השניה (שתי תוצאות לכל ווריאציה) התוצאה התקלקלה, ובהרבה. וכן קורה **בכל** 68 התוצאות לרשימה הראשונה. תוצאה כוללת של 136 קלקולים מול 0 שיפורים, נראית **מאד** לא סבירה גם לפי התזה של MBBK. האומנם רק ציפיות נאיביות לגבי התוצאות (ראה כיצד הם מנתחים תוצאות אלו בעמ' 169) מנעו בעד MBBK להבחין כי התוצאה הכוללת "יותר מדי טובה" לפי השערת המחקר שלהם?

הפתרון לכך הוא פשוט ומדהים:

(1). המעבר לחזקה ראשונה הוא הדומיננטי. ולכן לצורת הפונקציה הריבועית, ממנו לוקחים שורש ריבועי, תהיה רק השפעה **משנית** על התוצאה:

התבוננות בתוצאות בטור הימני לעומת המקבילות להן בטור המרכזי מגלה שיש תמיד קלקול גדול בתוצאה. גאנז [11] השווה את אוכלוסיית התוצאות עבור הפונקציות הריבועיות לעומת אוכלוסיית התוצאות עבור הווריאציות מן החזקה הראשונה. הוא עשה זאת עבור P4 (העמודה השלישית משמאל בכל טור), והוא מדווח:

> "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."

מאוחר יותר השווה גאנז [12] **גם את יתר שלוש העמודות, והרי הסיכום שלו:**

> "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). יתר על כן, **צפוי מראש שהרכבת** שתי הווריאציות - השינוי לפונקציה ריבועית מסוג המקלקל את התוצאה, בווריאציה שהיא המעבר לחזקה ראשונה (המקלקלת אף היא את התוצאה) - רק תהרוס **עוד יותר** את התוצאה. MBBK **עצמם כותבים בעמ' 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."(Emphasis ours).

(3). MBBK ידעו זאת מראש, **לפני** הבדיקה על כל 34 הפונקציות, ולא רק באופן תיאורטי. מתוך בדיקות שערך מקי לפני כשלוש שנים [13] על שלוש פונקציות בחזקה ראשונה, הוא ידע שהוצאת השורש הריבועי הורסת את התוצאה. יתר על כן, לאחר שפירסם יחד עם בר-הלל ובר-נתן את התוצאות עבור ארבע פונקציות כאלה ב- CHANCE [4], כתבנו במפורש באותו עיתון [14] שאנו מצפים מראש לכך שווריאציה בחזקה ראשונה תהרוס את התוצאה.

(4). נובע מכאן, שכל 34 הווריאציות בחזקה הראשונה הן בעצם אותה "אותה הגברת בשינוי אדרת". לכן, כל 68 התוצאות עבור הרשימה השניה הן בעצם רק **שתי** תוצאות, וכן הדבר לגבי 68 התוצאות עבור הרשימה הראשונה. אסור היה להם בשום אופן לקחת בעצם אותה וריאציה 34 פעמים, ולהציג זאת כ- 34 וריאציות.

הצגת הדברים כאילו יש כאן 2X68 תוצאות שליליות היא הטעיה חמורה.
הטעיה זו מטילה צל כבד על כל הווריאציות.

לקורא שאינו סטטיסטיקאי ננסה להמחיש זאת במשל. נניח שגלילאו משתמש לראשונה בטלסקופ ומגלה שישנם 4 ירחים לכוכב הלכת "צדק". כמובן, התגלית בניגוד למוסכמות, והמתנגדים חושדים בו ברמאות. המתנגדים מנסים לעשות ווריאציות על הניסוי שלו ולראות אם מקבלים אותה תגלית. למשל, הם מחליפים את עדשות הזכוכית בטלסקופ בעדשה אחרת, ומנסים לצפות לעבר ה"צדק". התוצאה: לא נצפים שום ירחים. גלילאו מתמרמר: הוא השתמש בעדשה מרכזת לאוביקטיב של הטלסקופ, ואילו המתנגדים השתמשו בעדשה מפזרת. הבחירה שלו בעדשה מרכזת נבעה מעצם הניסוי ולא היתה סתמית, ואילו "הווריאציה" שנעשתה חורגת

מכך לגמרי. [**אצלנו: עצם השימוש בחזקה הראשונה עבור פונקציה** *delta* **הוא שגוי**]. המתנגדים טוענים שגלילאו לא הצהיר מראש שעדשה מפזרת היא אמצעי שגוי. [**אצלנו: כך טענו מקי ושות' ב-** CHANCE]. ואז מבצעים המתנגדים סידרה של 33 ווריאציות עם 33 עדשות מפזרות עבור האובייקטיב, כל עדשה בעלת פוקוס אחר.
כמובן, שלא נצפים שום ירחים. המתנגדים טוענים שלא יתכן שגלילאו היה כה בר מזל והצליח לצפות בירחים, בעוד שהם עשו זאת ב- 34 ווריאציות ולא גילו דבר. [**אצלנו: MBBK טוענים שהם עשו 34 ווריאציות, וקיבלו תמיד תוצאה גרועה יותר**].

ההטעיה ברורה. המתנגדים בדקו בעצם וריאציה אחת: עדשה מפזרת. שים לב לכך, שאפילו אם הוריאציה זו היתה מוצדקת, אסור היה להם לחזור עליה עוד 33 פעמים ולספור זאת כ- 33 ווריאציות נוספות!

נעשה חסד עם MBBK, **ונמחוק** לפחות את 33 הווריאציות **שנוספו** בחזקה הראשונה.
אסור להביא אותן בחשבון בשום חישוב.

(ב) העלמת עיקרון המינימליות.

(1). המחקר שעשינו מבוסס על שני עקרונות מרכזיים. אחד מהם הוא עקרון המינימליות. עקרון זה מודגש מאד בכל הפירסומים שלנו המתארים את התופעה. כבר בפרה-פרינט הראשון [15] כתבנו:

> "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)

ולמשל, במאמרנו בסטטיסטיקל סאיינס [1] כתבנו:
"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)
כלומר הדגשנו היטב את מרכזיות עקרון המינימליות.
בעקרון המינימליות שני מרכיבים:
(א). אנו טוענים שהתופעה נשענת על אותם ELSs שהם מינימליים בקטעים גדולים בטקסט.
(ב). בחישובים אנו מקנים משקל (weight) גדול יותר לאותם ELSs שהם "יותר" מינימליים, דהיינו, שהם מינימליים על קטעים גדולים יותר.

(2). והנה, הקורא את מאמרם של MBBK אינו יכול אפילו **לנחש** שיש עקרון מרכזי של מינימליות. במבוא למאמרם (עמ' 151) הם מתארים את התופעה ומזכירים מפגש של מלים סתם, מבלי להזכיר כלל את היסוד: הופעות מינימליות. אומנם, בניספח A, בו יש תיאור מתמטי של התופעה, הם נאלצים להקדיש כמה שורות (בעמ' 168) כדי להגדיר בצורה פורמלית את המושגים "domain of minimality" ו- "domain of simultaneos minimality, ללא שום מלת הסבר לקורא מאין צצו פתאום מושגים אלה, שלא הוזכרו קודם (שלא לדבר על כך שלא הוזכר שזה עקרון יסודי). מושגים אלה גם אינם מוזכרים עוד במאמר עצמו. גם במקום שהם מצביעים על האספקטים המרכזיים של התופעה, לא מוזכר כלל עקרון המינימום:

> "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)

יש בדברים אלה התעלמות ברורה מעקרון המינימליות. הדברים בולטים ביותר לאור העובדה כי מקי עצמו ציטט את הקטע הבא ממאמרנו בדו"ח שלו [5], בהקשר לווריאציות שלו:

> "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". (WRR1, Pg. 431, emphasis ours).

הקורא יכול לתהות, אולי התעלמות זו אינה מכוונת אלא תוצאה של רשלנות. אולם להלן נראה כי התעלמות זו שמשה כרקע ליצירה מכוונת של וריאציות הסותרות את עקרון המינימליות.

(3). כאשר אנו מגיעים לנספח C, אנו מוצאים שם כי שתי טבלאות (מתוך שש) של וריאציות, טבלאות 8-7 עוסקות בווריאציות הקשורות קשר אמיץ עם עקרון המינימום.

הקורא במאמרם בלבד, בוודאי לא יכול להבין מה מרכזי עיקרון המינימליות, ולכן לא יחשוד שווריאציה "תמימה" (מס' 2 בטבלה 7) לא רק שאינה "natural" או "minor", זו פשוט התעלמות ממאפיין מרכזי של התופעה! אומנם, כאן הם השתמשו בהופעות המינימליות [לפי מרכיב (א) של עקרון המינימליות (ראה לעיל סעיף (1)], אך לא נתנו להן שום משקל מיוחד [בניגוד לנדרש במרכיב (ב) של עקרון המינימליות].

בניסויים הבאים הם הגדילו לעשות. הם פשוט "חתכו" וזרקו את ההופעות המינימליות עצמן. כך נעשה ב- 4 הווריאציות בסוף טבלה 8, בהם "חותכים" מן הנתונים דווקא את הדילוגים הקצרים ביותר שהם חלק מן ההופעות המינימליות, וזורקים אותם לאשפה. וכך קרה בניסוי "הפשוט" הבא:

> "…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." (Pg. 171, emphasis ours).

אבל בתחום הנזכר בין 10 ל- 20 נשארו מעט מאד הופעות שהן עדיין "מינימליות בקטעים גדולים בספר". לכן לפי השערת המחקר שלנו **אנחנו** מצפים לכשלון. הם פשוט "חתכו" וזרקו כמעט את כל ההופעות המינימליות בקטעים גדולים. דומה הדבר למי שמצא בצפייה בטלסקופ את ירחי ה"צדק", ובאו המתנגדים סובבו את הטלסקופ ב- 180 מעלות, והם מצביעים על כך שלא רואים את הממצאים! הקורא במאמרם אינו יכול לנחש שזו דווקא ראיה **לנכונות** השערת המחקר של WRR: שכן בזה הוכיחו MBBK שהתופעה אכן נשענת דווקא "על ההופעות המינימליות בקטעים גדולים".

אלו רק דוגמאות. בעניין טבלה 8 ראה עוד בסעיף הבא.

(ג) נשוב לטבלה 8 במאמרם, לארבע הווריאציות האחרונות בהן הם חתכו וזרקו את ה-ELSs שהם בדילוג פחות מ- 3, 4, 5, או 10. בסעיף הקודם הצבענו על כך, שווריאציות אלו צפויות **מראש** להיות הרסניות:

(1). כיוון שהתופעה נשענת על ההופעות המינימליות.

(2). בדילוגים קצרים אלה צפויות להיות עיקר ההופעות המינימליות של כמה וכמה ביטויים. אבל מתברר שאין זה הכל. מצאנו במאמרם כי הם בדקו ומצאו והדגישו כי

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

הכינוי המצליח עליו כתבו כאן הוא "הראב"י", והוא מצליח במידגם השני. ההצלחה שלו נובעת מהופעותיו בדילוג 2. מקי יודע זאת היטב, כי כבר בדו"ח הראשון שלו [13], **לפני** שהוצעו וריאציות כאלה, הוא חיפש תירוצים (אחרים) כדי להיפטר מן המפגשים המוצלחים של "הראב"י" בדילוג 2.

מתברר שהוא מצא דרך לעשות זאת על ידי הווריאציות הנ"ל, כאשר מובטח לו **מראש** שהתוצאה במידגם השני תתקלקל! (אגב, כך מובטח לו גם הרס עוד מפגשים מוצלחים כגון "מהרי"ל – בכ"ב אלול" ו"המלאך – בי"ב תשרי").

ראוי להדגיש: לפי השערת המחקר אנו מצפים להצלחה של ההופעות המינימליות של "הראב"י" עם תאריכיו, וההופעות המינימליות העיקריות שלו צפויות מראש להיות בדילוג 2 (זאת לפי אורך המלה ושכיחויות האותיות המרכיבות אותה).

(ד) מתברר, שהמודל האמיתי של MBBK לעניין "מחקר הווריאציות" הוא של "מפסידים בלבד".

בטבלה 5 שלהם ישנן 33 ווריאציות ריבועיות כנזכר לעיל. והנה, נתקלו MBBK בבעיה: הסטטיסטיקה P4 (שאת השינוי בה הם מציגים) דווקא **השתפרה** 19 פעמים מתוך ה- 33. מה הם עושים? ראשית, מצרפים MBBK גם את 34 הווריאציות בחזקה הראשונה לחשבון כדי להוריד את אחוז ההצלחה ולהציג אותה כ- 19 מתוך 67. הראינו לעיל, בסעיף **(א)**, שזו הטעיה.

המצחיק הוא, ש-MBBK לא הסתפקו בכך. לפי הכתוב במאמרם בעמ' 169 היו להם כנראה ציפיות נאיביות לקבל 0 מתוך 67. לכן מביאים MBBK **תרוצים א-פוסטריוריים** כדי לבטל גם את התוצאה של 19 מתוך 67, תרוצים שנתפרסמו לראשונה במאמרם.

עיקר התירוץ שלהם הוא, ש"תפירת" רשימת הכינויים המיוחסת לנו היתה לא רק לצורך אופטימיזציה של P4 אלא גם לצורך אופטימיזציה אחרת, ולכן אי אפשר ללמוד דבר מהשתפרות P4. לא רק שסיפור האופטימיזציה הנוספת הוא שטות נוספת (ראה בסוף פרק ג), יש לפנינו גם טעות במתמטיקה וגם הטעייה חמורה.

1. טעות במתמטיקה: אופטימיזציה נוספת אינה צריכה למנוע מ-P4 המקורי להיראות כאופטימום.
2. והחמור מכל – הטעיה. נראה, ששיטתם הכוללת היא כזו:

- שאם הווריאציות יראו אופטימום בתוצאה המקורית – זה יוכיח ש-WRR עשו אופטימיזציה על הכינויים,
- ואם הווריאציות לא יראו אופטימום בתוצאה המקורית – זה לא אומר כלום, כי תמיד אפשר לבטל את משמעות התוצאה על ידי "סיפורים" א-פוסטריוריים.

המשטרה סוגרת מועדוני הימורים הפועלים לפי מודלים כאלה.

(ה) בפרק א סעיף 9 הצבענו על הליקוי היסודי הבא: הבחירה שלהם בארבע סטטיסטיקות מסוימות כדי להציג את תוצאות הווריאציות (הקטע בשם "What measures should we compare" בעמ' 160), היא בחירה א-פוסטריורית של חלק מן המדידות. הם בחרו רביעיה אחת מתוך מספר עצום של צרופים אפשריים, על סמך סיפורים א-פוסטריוריים. זה ליקוי עקרוני.

אנו נייחד פרק שלם, פרק ג, כדי לפרוש את התוצאות המלאות עבור הסטטיסטיקות השונות, ונראה שאומנם הבחירה של MBBK מסלפת באופן חמור את התמונה העולה מן הניסויים. כאן נסתפק בדוגמא.

בהצגה מסולפת זו של הווריאציות הם הגיעו לשיא הגיחוך במקרה של הווריאציות על "Cut-off defining P1" (טבלה 10). לפי רביעיית הסטטיסטיקות שבחרו להציג יוצא שהם מראים תוצאות על P2 (ו- P4 שהוא אנלוג שלו), בניסוי המיועד לבדוק ווריאציות על P1! שהרי ווריאציות אלה נועדו להשפיע על P1 (ו- P3 שהוא אנלוג שלו), ובעקיפין על r1 ו- r3.

טבלה 10 נראית לפי בחירתם כך:

	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

טבלה 1

[הוספנו כאן את הסימון הבא: L1=הרשימה הראשונה, L2=הרשימה השניה.]

אבל אם נציג את התוצאות עבור הסטטיסטיקות הרלוונטיות, התמונה תהיה כדלקמן:

עבור L1:

Cut-off defining P1	P1	r_1	P3	r_3
0.05	475487	11.3	134	3.2
0.1	386357	55.7	1207	31.1
0.15	2639	15.9	74	5.8
0.2 (WRR)	1	1	1	1
0.25	0.0024	0.08	0.019	0.1
0.33	0.0008	0.043	2.5	5.7
0.4	0.001	0.09	0.63	3.7
0.5	0.00013	0.016	0.018	0.32

טבלה 2

עבור L2:

Cut-off defining P1	P1	r_1	P3	r_3
0.05	105048	19.0	5157	8.5
0.1	133	1.9	6.6	0.24
0.15	145	4.0	14.4	1.3
0.2 (WRR)	1	1	1	1
0.25	0.00032	0.014	0.000015	0.0019
0.33	0.00034	0.052	0.00011	0.016
0.4	0.0083	0.19	0.0048	0.12
0.5	0.055	0.85	0.05	1.0

טבלה 3

חשוב להעיר, כי P1 ו-P2 שימשו כסטטיסטיקות היחידות להערכת הצלחת L1 ו- L2. לכן, אם נעשתה אופטימיזציה, הרי נעשתה ביחס ל- P1, או ביחס ל- P2, או – מה שסביר יותר - ביחס ל- Min(P1-P2). נציג את התוצאות לפי הסטטיסטיקה 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	0.0069
0.33	0.0008	0.0073
0.4	0.001	0.18
0.5	0.00013	1.0

טבלה 4

בטבלה זו ישנו מידע חשוב ביותר. הרי MBBK מטילים חשד באמינות של WRR על ידי נסיון להביא ראיות עקיפות לכך, ש- WRR עשו כל מניפולציה כדי להשיג תוצאות מרשימות יותר. והנה, מן הווריאציות בטבלה זו, **מתברר שישנה ראייה ישירה, חזקה וברורה לכך, ש- WRR פעלו ביושר גמור**. אין דבר קל ופשוט מבחירה א-פוסטריורית של ה-"Cut-off defining P1". הצצה בטבלה מגלה, כי WRR יכלו לשפר את התוצאה שלהם (שנמדדה על ידי ערכי P1 ו-P2) **אלפי מונים** על ידי בחירה ב- Cut-off מתאים.

אבל הקורא המעיין בטבלה 10 שלהם, אינו יכול להגיע לכל המידע החשוב זה. גם אם יתקל במה שכתבו MBBK בעניין זה:

> "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)

עדיין אינו יכול לנחש מדבריהם את הנתונים החשובים שבטבלאות 2-4 דלעיל, ובוודאי שלא יוכל להסיק מדבריהם שכאן ישנה ראייה חזקה לכך ש- WRR פעלו ביושר.

דוגמא בולטת זו של הטעיה באמצעות צורת ההצגה היא, לדעתנו, רק דוגמא אחת להטעיה שיטתית באמצעות צורת ההצגה. נדון בכך בהרחבה בראש פרק ג.

(ו) האמת היא, שאין אנו צריכים לחשוד בMBBK שבחרו תמיד את ההצגה הגרועה ביותר מבחינתנו. למעשה, התהליך יכול היה להיות גם הפוך: כיוון שלפי התיזה שלהם הם צריכים להראות אופטימיזציה בעיקר לגבי סטטיסטיקות r (הדירוגים במבחן הפרמוטציות), הם יכלו לחפש (או ליצור, ראה סעיפים **(א)**,**(ב)**, ו**(ג)** לעיל) דווקא את הווריאציות שמקלקלות במיוחד סטטיסטיקות אלו. וכיוון שאוסף הווריאציות לא רק שאינו סגור, אלא הוא פרוץ לחלוטין, אפשר לחפש (או ליצור) וריאציות כאלה בנקל.

למשל, מתוך עיון בווריאציות בטבלאות 5-10 ניתן לאבחן קבוצה A של תוצאות עבור וריאציות מן הסוג הבא: וריאציות שהן נקודות דגימה עבור פרמטרים או ספים שהיו אפשריים בניסוי המקורי. סך הכל מטפלים MBBK ב- 7 ספים או פרמטרים כאלה. כיוון שיש כמה נקודות דגימה עבור כל פרמטר או סף, ישנן בסך הכל 44 נקודות דגימה, המחולקות ל- 7 מחלקות. לכל נקודת דגימה קוראים MBBK "וריאציה".

חישוב הווריאציות נעשה בשני שלבים (לפחות). הדו"ח של מקי [5] כלל תת-קבוצה A1 של תוצאות מתוך A, ובה תוצאות עבור 22 וריאציות (עבור L1 היו תוצאות רק עבור 20 מתוך 22 הווריאציות הללו). שאר התוצאות נמדדו בשלב מאוחר יותר, ונסמן תת-קבוצה זו ב- A2=A-A1. בכל שלב, בחירת נקודות הדגימה נעשתה באופן שרירותי.

השיטה של בחירה שרירותית של נקודות דגימה עבור פרמטר מסויים **מאפשרת** הטעייה. יש לכך שני היבטים:

(1). **<u>הכפלה מכוונת</u>**: בסעיף א לעיל, הראינו כיצד הכפילו MBBK את מספר הווריאציות בטבלה 5, תוך מגמה שאחוז התוצאות המצביעות על קילקול לא ירד. הם עשו זאת על ידי הכנסת 33 הווריאציות בחזקה הראשונה (לפרטים ראה שם). במקרה שלפנינו הם פעלו באותה מגמה: להכפיל את התוצאות עבור נקודות הדגימה שהיו בידם בשלב הראשון (A1) על ידי הוספת תוצאות נוספות (A2), כך שאחוז הקילקולים לא ירד בשלב השני.

הטכניקה לכך היתה פשוטה: הם דגמו שוב ושוב מתוך **אותן** 7 מחלקות. כלומר, לאחר שדגמו פעם ראשונה ממחלקה מסויימת, והגיעו לכך שרוב התוצאות מצביעות על קילקול ורק מיעוט על שיפור, חזרו ודגמו מאותה מחלקה, למרות שהתוצאות כבר ידועות (בערך) מראש.

אבל, אם נשווה את תוצאות הדגימה בשני השלבים, נראה כי למעשה, התוצאות היו עוד יותר תואמות את רצונם של MBBK לאפשר "כיסוי" להצהרה [4] כמו זו:

> "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".

אם נשווה את סך כל התוצאות עבור ארבע הסטטיסטיקות של MBBK (שיפורטו בסעיף הבא) בשני השלבים, נקבל את התמונה הבאה:

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

טבלה 5

"better" – אלו המקרים בהם השתפרה התוצאה עקב הווריאציה.

"equal" – אלו המקרים בהם התוצאה לא השתנתה עקב הווריאציה.
"worse" – אלו המקרים בהם התוצאה התקלקלה עקב הווריאציה.

הרי לפנינו **עליה מרשימה בכמות הקילקולים ב-**A2: **43%!**
כיצד קרה הנס הזה, שהתוצאות של A2 השתנו ביחס כזה לעומת התוצאות של A1?

ננסה לברר האם גם ישנם עקבות לציפיות (או לכוונות) של MBBK שעיקר הקילקולים הם בסטטיסטיקות r (ראה סוף הציטטה לעיל). נגדיר את "יעילות ההרס" של הדגימה, $e(\mathrm{Ai})$, כמספר הקילקולים הממוצע לדגימה בקבוצה Ai:

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

טבלה 5א

שים לב לעלייה הדרמטית ביעילות ההרס ב- A2 דווקא בסטטיסטיקות r אותן מעדיפים MBBK!

שינויים כאלה לרעה, אינם יכולים לנבוע מהכפלה בלבד. וזה מחייב אותנו לבחון האם לא היה כאן גורם נוסף. כך אנו מגיעים לנושא של דגימה סלקטיבית.

(2). <u>דגימה סלקטיבית</u>: לא רק שאין שום שיטה קבועה וא-פריורית לבחירת הנקודות, ובכל פעם נלקחות נקודות הדגימה בצורה שרירותית. אלא מתברר שהרבה פעמים אפשר לדעת <u>מראש</u> ממגמת התוצאה עבור נקודות ידועות, מה תהיה התוצאה עבור נקודות אחרות, ואפשר לנצל זאת לדגימה סלקטיבית.

להלן נבחן בפרוטרוט את האפקט של ההכפלה ושל הדגימה הסלקטיבית עבור כל אחת מ- 7 המחלקות בקבוצה A.

(א). בטבלה 6 במאמרם, מציגים MBBK 7 נקודות מדידה עבור "value of *i*", בערכים 1, 2, 5, 15, 20, 25 ו- 50. נעתיק זאת כאן:

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	<u>0.6</u>	<u>0.3</u>	<u>0.2</u>
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

טבלה 6

אבל בדו"ח של מקי [5] בו מתואר ניסוי זה בראשונה, נמדדו רק הנקודות 2, 5, 15, ו- 20 (מודגשות ברקע אפור). נקודות 1, 25 ו- 50 נוספו רק **עכשיו**. מתוך סדרת התוצאות עבור 10, 15 ו- 20 אפשר לראות מגמה ברורה של קילקול התוצאות. אפשר היה לשער שדגימה בנקודות 25 ו- 50 תניב אף היא קילקול בתוצאות. ואכן, זה מה שקרה. גם השוואת התוצאות עבור הנקודות 5 ו- 2 מאפשרת לנחש שעבור הערך 1 נקבל קילקול התוצאות. ואכן, זה מה שקרה. MBBK דגמו דווקא נקודות אלה.

לעומת זאת הם לא דגמו כלל בין נקודות 5 ל- 15: זהו תחום בו **לא** מובטח מראש קילקול התוצאה. וזאת, למרות עמדתם המוצהרת (ראה לעיל פרק **א** סעיף 5) כי הם בודקים

דווקא "minor variations", "minimal changes", "small changes". הבה נעשה זאת במקומם:

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

טבלה 7

הם לא דגמו גם בין הנקודות 2 ל- 5, וגם לגבי תחום זה לא מובטח מראש קילקול התוצאה על פי המדידות בשלב הראשון:

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

טבלה 8

התוצאה החשובה ביותר ל-MBBK היא min(r1-r4) עבור L2 (התוצאה הימנית בכל רביעית תוצאות בטבלה), כי זו התוצאה של WRR. בפרק ה נכתוב על כך בהרחבה. כאן רק נשים לב לכך, כי בדגימות שהוספנו בטבלאות 7-8 משתפרת סטטיסטיקה זו 8 פעמים, ומתקלקלת רק פעם אחת מתוך סך כל 10 הדגימות. השווה נא תוצאה זו לכך, שלפי תוצאות דגימות MBBK משתפרת סטטיסטיקה זו רק 4 פעמים מתוך 135 הווריאציות שבטבלאות 5-10 במאמרם.

(ב). בטבלה 8 במאמרם אנו מוצאים 9 נקודות מדידה עבור "Expected ELS count of", בערכים 2, 5, 15, 20, 25, 30, 50, 75, ו- 100. נעתיק זאת כאן:

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

טבלה 9

אבל בדו"ח של מקי הוצגו תוצאות רק עבור הנקודות 5 ו- 15 לגבי L1, ועבור הנקודות 5, 15, 20 ו- 30 לגבי L2 (נמסר שם שהמדידות לגבי נקודות 20 ו- 30 עבור L1 לא נסתיימו, וכי יש בדעת מקי למדוד את 50 בעתיד). תוצאות אלה הדגשנו ברקע אפור.
מתברר, כי לאחר שנראתה מתוצאות אלה מגמה ברורה לקילקול התוצאה עבור ערכים קטנים מ- 5 וגדולים מ- 30, הוסיפו MBBK **מאוחר יותר** גם את התוצאות עבור 2, 50, 75, ו- 100. ויותר מזה, הם הרשו לעצמם להוסיף עוד נקודת מדידה, 25, בין שתי תוצאות ידועות (של קילקול התוצאה) עבור 20 ו- 30. כלומר, רוב הנקודות נוספו **לאחר** שהיתה ידיעה מוקדמת על המגמה הצפויה.
[הערה: עבור נקודות 2 ו- 5 צפויות התוצאות מראש להתקלקל בגלל שמספר המתחרים בדילוגים שאינם שווים צפוי מראש להיות קטן מהמקורי (אומנם מספר זה אינו ידוע במדויק מראש). האפקט של שינוי מספר המתחרים כשלעצמו גרם לקילקול התוצאה בפקטור של 1.69 ו- 1.36 בהתאמה בגלל גורם זה בלבד. הסבר על כך תמצא בסעיף (ה) להלן. אומנם, במקרה הנוכחי ישנם גורמים נוספים הגורמים לקילקול נוסף.]

גם כאן הם לא דגמו כלל בין נקודות 5 ל- 15: זהו תחום בו **לא** מובטח מראש קילקול התוצאה. וזאת, למרות עמדתם המוצהרת (ראה לעיל פרק **א** סעיף 5) כי הם בודקים דווקא "minor variations", "minimal changes", "small changes". **הבה נעשה זאת במקומם:**

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

טבלה 10

מעניין, אך גם כאן משתפרת הסטטיסטיקה min(r1-r4) עבור L2 פעמים אחדות. בייחוד אם נשוב ונזכור שסך כל ההשתפרויות בסטטיסטיקה זו בווריאציות ש MBBK הציגו בטבלאות 5-10 במאמרם הוא רק 4 מתוך 135.

(ג). בסעיף (**ה**) לעיל הבאנו את תוצאות הווריאציות מטבלה 10 של MBBK, והראינו מה רב הסילוף בצורת הצגת הנתונים. הבה נתבונן שוב בטבלה:

	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

טבלה 11

אנו מבחינים כי עד הערך של 0.2, דוגמים MBBK כל 0.05, ואילו בין 0.2 ל- 0.5 הדגימה שלהם דלילה יותר. האם יש לכך קשר עם העובדה כי התחום המניב השתפרויות הוא דווקא בקטע [0.2,0.5), כפי שניתן לצפות מן ההיסטוגרמות שפירסמו WRR? הבה ונדגום כל 0.05 גם בקטע זה:

	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

טבלה 12

אנו רואים שקיבלנו עוד שתי השתפרויות עבור Min(r1-r4) ב-L1: זה יבול לא רע בהתחשב בעובדה ש- MBBK "הקציבו" לסטטיסטיקה זו רק 13 השתפרויות עבור כל 135 הווריאציות. ואם מציגים את התוצאות בדרך הנכונה, דהיינו לפי הסטטיסטיקה Min(P1-P2), מקבלים השתפרות עבור כל הנקודות שהוספנו: 0.3, 0.35 ו- 0.45, גם ב- L1 וגם ב- L2.

[הערה: על רקע הניסוי הקטן שערכנו זה עתה, יובנו השינויים שעשו MBBK בנתונים אלה לעומת הדו"ח של מקי. שם, היו בקטע [0, 0.2) ארבע נקודות דגימה: 0.01, 0.02, 0.05 ו- 0.1, ואילו בקטע [0.2,0.5) שהוא גדול פי 1.5 היו רק שלוש נקודות דגימה: 0.25, 0.33 ו- 0.5. לדעתנו, ניסו MBBK ליצור רושם בהצגה הנוכחית, שהדגימה מפוזרת באופן יותר אחיד, על ידי הורדת הנקודות 0.01 ו- 0.02 והוספה של 0.15 ו- 0.4. כך השיגו רושם יותר מאוזן מבלי לשלם מחיר: השינוי נעשה כך שלא השתנה מאזן התוצאות, ועדיין הדגימה דלילה יחסית בתחום הצפוי להשתפר. בכל מקרה, שים לב למשחק החופשי של MBBK בנקודות הדגימה].

(ד). דוגמא נוספת לדגימה צפופה בתחום שבו צפויה התוצאה להתקלקל.

הערכת המפגשים בין הביטוי w לביטוי w' בדילוגים שווים נעשית על ידי השוואת המפגשים בדילוגים השווים, למפגשים של אותם ביטויים בדילוגים שאינם שווים: נערכת "תחרות" בין קבוצה של דילוגים לא שווים לבין הדילוג השווה על המפגשים ה"מוצלחים" ביותר. תוצאת ה"תחרות" היא הפונקציה $c(w, w')$, והיא שבר פשוט: a/m, כאשר

a= הדירוג של המפגש בדילוגים שווים, ו- m= סך כל המתחרים. אם a/m קרוב ל- 0, פרושו של דבר הצלחה- הדילוג השווה דורג באחד המקומות הראשונים. אם a/m קרוב ל- 1 פרושו כשלון - הדילוג השווה דורג בסוף.

בניסוי המקורי לא הובאו בחשבון אותם ערכי מפגשים שבהם היו מעט מדי "מתחרים" – פחות מ- 10. היתה לכך סיבה ברורה. תאר לעצמך שקיים מיפגש מוצלח ביותר בדילוגים שווים בספר בראשית: מיפגש בו ה- ELSs הנפגשים הם לא רק קרובים מאד זה לזה, אלא גם "נדירים" – הסיכוי להופעתם כ- ELS במקרה הוא נמוך מאד. בגלל סיכוי נמוך זה, לא יימצאו "מתחרים" בדילוג שאינו שווה: הם פשוט לא יופיעו! ה- ELSs ישתתפו לבדם בתחרות, ותוצאת התחרות תהיה 1/1 (כי מספר המתחרים הצטמצם לאחד). תוצאה זו מורה על כישלון גמור! (זכור, ככל שהערך של c קרוב ל- 0 זו הצלחה, וככל שהוא קרוב ל- 1 זה כשלון). גם אם יהיה עוד מתחרה (בדילוגים לא שווים), עדיין יהיה עיוות בלתי נסבל של התוצאה: 1/2 הוא ערך שאינו מצביע על הצלחה. כדי למנוע עיוות זה קבענו סף של m=10 מתחרים.

MBBK מציגים בטבלה 10 במאמרם, ערכים אחדים עבור סף זה (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

טבלה 13

כאן נדגמו נקודות עבור "Denominator bound" בערכים 2, 3, 4, 5, 15, 20 ו- 25. אולם לפי דו"ח מקי נדגמו קודם לכן רק הנקודות 2, 5, 15, ו- 20 (הדגשנו אותן ברקע אפור, והוספנו את ערך הסף של WRR להשוואה).

לפי האמור לעיל ידוע **מראש** שסף של 2 או 3 יכול רק להשפיע לרעה על ערכי P1-P4 (וסף של 4 על P1 ו- P3, כי עבורם כל תוצאה גדולה מ- 1/5 היא כשלון), ובעקבותיהם על r1-r4. ראוי לציון, שדווקא בטווח הספים הצפוי לקלקל, הם דוגמים בצפיפות: גם 2 וגם 3 וגם 4! הרי לפנינו עוד דוגמא כיצד דוגמים בצפיפות בתחום שבו צפויה התוצאה להתקלקל.

גם הדגימה לגבי הערך של 25 מאלפת. כל מי שקרא את רשימת ערכי *c(w, w')* עבור L2 בפרה-פרינט שלנו משנת 88 (ו-MBBK קראו בו) יודע, כי מכל 163 הזוגות ב- L2, רק במקרה **אחד** היה המכנה פחות מ- 20: היה זוג אחד עבורו התוצאה היתה 4/19, וכי לא היה אף זוג בעל מכנה של 20, 21, 22, 23, 24 או 25. כאשר MBBK הציבו סף של 20, הם פשוט מחקו את התוצאה של 4/19, ובזה קלקלו את התוצאה כפי שהוצגה על ידם. עכשיו הוסיפו MBBK **עוד דגימה**: סף של 25, העושה בדיוק מה שעשה הסף של 20: מוחק את התוצאה של 4/19!

(ה). דוגמא נוספת לדגימה צפופה בתחום שבו צפויה התוצאה להתקלקל.

כאמור בסעיף הקודם, חישוב הפונקציה *c(w, w')* המקורית נעשה על ידי "תחרות" בין הדילוג השווה לדילוגים שאינם שווים. סך הכל היו 125 מתחרים. נניח שהדילוג השווה היה המצטיין בתחרות עבור זוג ביטויים מסויים. התוצאה במקרה זה היא $c=1/125$. שאלה: מה יקרה אם במקום 125 מתחרים יהיו רק 25 מתחרים? תשובה: התוצאה תהיה 1/25, כלומר, התוצאה תהיה **גרועה** פי 5. גם אם מספר המתחרים יהיה 49 או 81 התוצאות צפויות **מראש** להתקלקל (במקום 1/125 נקבל 1/49 או 1/81). ההשפעה תהיה על הסטטיסטיקות P2 ו- P4 ועל r2 ו- r4. אנו זוכרים היטב, ש-MBBK (בעמ' 155) הסבירו כי השפעת הערכים **הקטנים** ביותר על הסטטיסטיקות הללו היא רבה.

יש ב- WRR חמישה ערכי 1/125. אפילו אם נקפיא את כל הערכים האחרים, ונשנה רק חמישה ערכים אלה כך שבמקום 125 מתחרים ניקח 25, 49 או 81 מתחרים, הקילקול ב- P4 עבור WRR יהיה בפקטור של 19.76, 5.82 ו- 2.28 בהתאמה!

איך קובעים את מספר המתחרים?

כבר מבואר במאמרנו המקורי כי המתחרים בדילוגים שאינם שווים נוצרים על ידי שיבושים (perturbations) לדילוג השווה באמצעות 3 משבשים: (x,y,z). כל אחד מהם מן המשבשים יכול לקבל אחד מ- 5 ערכים: 2-, 1-, 0, 1, ו- 2. לכן ישנם סך הכל 5x5x5=125 אפשרויות לשבש את הדילוג השווה, וזה גם מספר המתחרים.

אפשר לשנות את מספר המתחרים בשתי דרכים:

- אפשר לשנות את טווח הערכים *2n+1* שיכול לקבל כל משבש. אם $n=3$ נקבל 7x7x7=343 מתחרים. עבור $n=4$ נקבל 729 מתחרים, ועבור $n=5$ נקבל 1331 מתחרים. ולהיפך: אם נקטין את *n* ל- 1, נקבל רק 3x3x3=27 מתחרים.
- אפשר לשנות את מספר המשבשים. למשל, במקום שלושה משבשים (x,y,z) אפשר לקחת רק זוג משבשים (x,y). ואז, אם $n=2$, יש לנו 5x5=25 מתחרים, אם $n=3$ ישנם

49 מתחרים, ואם $n=4$ ישנם 81 מתחרים.

כשאנו מצוידים בכל הידיעות הללו נתבונן בטבלה 9 של MBBK. בה הם דוגמים נקודות עבור שינויים אפשריים במספר המתחרים:

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

טבלה 14

הדגימה המסומנת ברקע אפור היא עבור $n=3$ ושלושה משבשים, כלומר 343 מתחרים. היא מופיעה כבר בדו"ח של מקי. אבל ישנה בדו"ח של מקי עוד נקודת דגימה: עבור $n=1$ ושלושה משבשים, כלומר 27 מתחרים. מן התוצאות עבורה אפשר ללמוד כי אכן, כצפוי, זו וריאציה המקלקלת את התוצאות באופן בולט. כאשר באו MBBK להוסיף דגימות בשלב הבא, עמדה לפניהם רק האפשרות להוסיף את המקרים $n=4,5,6,\ldots$ המגדילים את n. אבל הם היו מעונינים דווקא שמספר המתחרים יהיה קטן מ- 125: כפי שהסברנו לעיל זה צריך לגרום לקילקול התוצאה. לכן, במקום להציג את המקרה $n=1$ הנמצא בדו"ח של מקי, הם פרטו אותו ל- 3 תוצאות בדרך הבאה: הם עברו מ- 3 משבשים לשני משבשים, ואז הם יכולים להציג תוצאות עבור 25, או 49, או 81 מתחרים. אלה שלוש הווריאציות האחרונות בטבלה 14.

לעומת החריצות וכושר ההמצאה שאפשרו את תוספת הנקודות עבורן יש פחות מ- 125 מתחרים, אנו רואים ש- MBBK התעייפו מהר מאד כאשר דגמו נקודות להן מספר מתחרים גדול יותר. הם הסתפקו בנקודה אחת ויחידה: $n=4$ ושלושה משבשים (שורה שניה בטבלה 14). חבל, אם היו ממשיכים לדגום היו מקבלים שיפורים משמעותיים לתוצאות:

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

טבלה 15

כאן הוספנו את $n=5,6$.

(ו). בטבלה 6 במאמרם, ישנן 4 נקודות דגימה עבור "Minimum row length", בערכים 3, 4, 5, ו- 10:

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

טבלה 16

אבל, בדו"ח של מקי נמדדו רק הנקודות 1 ו- 10, ונמסר ש- 5 לא נמדדה עדיין. כלומר, כאן הושמטה התוצאה עבור 1, ונוספו תוצאות עבור נקודות 3, 4 ו- 5. צריך להדגיש, שלא היה בניסוי WRR פרמטר כמו "Minimum row length", כפי שברור גם מן ההגדרה של $H(d,d')$ במאמרנו. לאחר הניסוי התברר, שהיה "באג" בתוכנה ולכן לא נמדדו טבלאות ובהן אות אחת

בשורה. מקי בדק זאת כבר בדו"ח שלו – זו נקודת הדגימה 1 – ונוכח לדעת כי באמת, עבור הנקודה 1 **אין** קילקול בתוצאה. הוא מודה שם כי לא נזכר במאמרנו סף שכזה.

אבל, למרות ש-MBBK נוכחו לדעת כי בוודאי לא היתה כאן קביעת סף מכוונת, בחרו להעלים מידע זה מן הקורא, ובתיאור שיטת המדידה שלנו בנספח A במאמרם, הם מציגים את הגדרת $H(d, d')$ (עמ' 167), כאילו ההגדרה כללה סף Minimum row length=2. וגם כאן, בהצגת נקודות הדגימה, הם השמיטו את התוצאה עבור הערך של 1.

(ז). המחלקה האחרונה שנדון בה, כוללת את 5 הוריאציות האחרונות בטבלה 8 במאמרם:

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

טבלה 17

בדו"ח של מקי נדגמו נקודות 3 ,4 ו- 10, שהדגשנו ברקע אפור. כאן נוספו שתי דגימות.

בסעיף **(ג)** הסברנו, שוריאציות בהן "חותכים" את ההופעות בדילוגים הקצרים צפויות **מראש** להיות הרסניות: וזה משותף לכל ארבע הנקודות האחרונות בטבלה, ללא הבדל באיזה שלב נדגמו.

עד כאן דברנו על שיטת ההכפלה ועל הדגימה המכוונת. לצורך הדיון, אף הנחנו כי הדגימה בשלב הראשון היתה בדרך כלל תקינה. אבל, מדוע להניח זאת? עצם המשחק החופשי בנקודות הדגימה, מעורר תהיות רבות.

עצם ההוספה **המאוחרת** של נקודות מדידה שנצפתה כאן, ובמיוחד כשניתן היה לחזות את התוצאות מראש, מעוררת את השאלה: **בכמה שלבים** נדגמו הנקודות בדו"ח **המקורי** של מקי?!

<u>לסיכום</u>: החופש הגמור בבחירת נקודות הדגימה, יחד עם הראיות הברורות לניצול חופש זה, הופך דגימה זו לחסרת כל ערך.

(ז) בפרק א סעיף 3 ציינו כי אם קיימת פלוקטואציה שגרמה לשיפור התוצאה בניסוי המקורי, הרי הפעלת הוריאציות יגרום בדרך כלל לקילקול התוצאה. והנה, במאמרם (בפרק 4) אכן מאריכים MBBK להסביר כי בניסוי של WRR היתה פלוקטואציה ששיפרה את התוצאה.

בכל ניסוי אפשר לאבחן באופן א-פוסטריורי כל מיני פלוקטואציות. מה ערכה של ביקורת מסוג זה? – על כך נדון במקום אחר.

אבל מבחינת העניין הנידון כאן, סומנה פלוקטואציה זאת על ידי מקי [13] **לפני** עריכת "מחקר הוריאציות", ולכן היה על MBBK לקחתה בחשבון, מבחינת הערכת תוצאות הוריאציות.

האם MBBK לא שמו לב לנקודה זו, והתייחסו לניסוי המקורי **כפי שהוא**, מבלי לגעת בנושא הפלוקטואציות? – לא נוכל לומר זאת. בפרק 3 במאמרם, הם מצביעים על פלוקטואציה נוספת: הפעם פלוקטואציה המקלקלת את תוצאת WRR. הכוונה לכך, שדיאקוניס הציע להשתמש ב- 1,000,000 פרמוטציות לצורך הניסוי הקובע, וכך התקבלה תוצאה של 4/1,000,000 עבור min(r1-r4). תוצאה זו היתה נוחה ל MBBK: <u>באותו הזמן</u> שפירסמו באינטרנט את נוסח מאמר ה- Stat. Sci., פירסמו מקי ובר-נתן <u>באותו אתר</u> את מאמרם על עבודתם ב"מלחמה ושלום"[16], ובו הם קובעים את תוצאת min(r1-r4) של WRR להיות 400/100,000,000 (וזאת כדי להשוות לתוצאה שלהם שהיא 57/100,000,000).

אבל כאן, לצורך "מחקר הווריאציות" MBBK טוענים שזו "שגיאת דגימה". הם מבטלים את הפלוקטואציה על ידי בחירתם כבסיס להשוואה לצורך "מחקר הווריאציות" לא בערך של 400/100,000,000 אלא בערך של 68/100,000,000.

חזרנו וביצענו את הווריאציות בתנאי הניסוי המקוריים, עם אותן 1,000,000 פרמוטציות. כך מצאנו, כי על ידי ביטול פלוקטואציה זו בלבד, שינו MBBK את תוצאות "מחקר הווריאציות" כדלקמן:

במקום: 11 שיפורים, 24 תיקו ו- 67 קילקולים,
יש להם עכשיו: 4 שיפורים, 13 תיקו ו- 85 קילקולים.

זו דוגמה לכך, כי פלוקטואציה שהשפעתה על התוצאה היא רק בפקטור 6, יכולה להשפיע בצורה חזקה על תוצאות "מחקר הווריאציות".

זו גם דוגמה לכך, כי MBBK השתמשו בכך לקדם את מטרתם.

2. <u>הטעויות</u>:

(א) בטבלה 6 הם החליפו את פונקצית μ כמה אופנים שגויים. השגיאה הבולטת ביותר היא החלפת ההגדרה המקורית $\mu = \delta^{-1}$ בהגדרה $\mu = -\delta^2$. להזכיר: התלות של δ במרחק r בין שני ELSs היא: $\delta \sim r^2$ ולכן $\mu \sim {}^{1}\!/_{r^2}$. לפי השינוי שלהם: $\mu \sim -r^4$. ברור שלפי הגדרה חדשה זו החלק הדומיננטי בהגדרת הקירבה בין שני ELSs יהיה של ה- ELSs המרוחקים, ולא של הקרובים!

המשל הבא יעזור להמחיש את האבסורד בשינוי זה. חוק של $1/r^2$ מאפשר לבדוק במבחנה השפעה מקומית של מולקולות של מגיב כימי מסוים, על אטום ברזל במולקולה של המוגלובין. ברור שהשפעת המולקולות והחלקיקים יורדת באופן דרסטי כאשר מתרחקים מן האטום הנדון. לכן אפשר בכלל ליצור כימיה ופיזיקה מקומית כדי לדון בבעיה כגון זו, מבלי לקחת בחשבון את השפעת המולקולות הרחוקות, למשל של המאדים.

נעשה שינוי "קטן" ("slight" - כדברי MBBK) לחוק הנ"ל, כך שהחוק החדש יראה כ- $-r^4$. עכשיו תהיה השפעת המאדים הדומיננטית, והשפעת המולקולות במבחנה תתבטל לעומתה!

רק הטיה קיצונית של MBBK גרמה להם לעשות שגיאה כה גסה. אפשר לנחש, כי מעולם לא הדפיסו בעיתון מדעי שגיאה שכזאת. דברינו בסעיף זה אמורים גם לגבי השינויים ל- $\mu = -\delta$ ול- $\mu = -\ln\delta$, הנמצאים באותה הטבלה.

(ב) בפרה-פרינט הראשון (86), אנו מדגישים את חשיבותם של <u>שני</u> מרכיבים במפגש הגיאומטרי בין שני ELSs: שכל אחד מן ה-ELSs יהיה ממוקד על פני הטבלה (או הגליל) הדו-ממדי, כלומר יהיה בעל "small localization parameter" (f קטן), ושהם יהיו קרובים זה לזה (l קטן). ראה עמ' 8-9 ו 29-30 שם. ואילו MBBK מתעלמים מכך בחלק ניכר מן הווריאציות המוצגות בטבלה 5.

(ג) התוצאה הסופית ב- WRR מתקבלת מהצטברות ערכי c עבור זוגות הביטויים. לכן **צפוי מראש** שקיצוץ במספר הזוגות יביא לירידה במובהקות הקבוצה הנשארת. דוגמא קיצונית: אם נוריד את מחצית הזוגות (באופן אקראי), אנו צפויים לקלקל את התוצאה פי 1000. ולכן צפוי הדבר, שווריאציות בהן מורידים את מספר הזוגות יקלקלו את התוצאה.

אבל MBBK עשו שש וריאציות לפחות בטבלאות שבנספח C, בהן היה צפוי קיצוץ במספר הזוגות:

- שתי הווריאציות הראשונות בטבלה 8 (בגלל הסף הנמוך לדילוג לא יופיעו כל הזוגות בדילוג שווה).
- הווריאציות 5-6 בטבלה 9, שם הורדת מספר המתחרים גרם לכך שלכמה זוגות לא היו מספיק (10) מתחרים, והם נמחקו.
- הווריאציות 6-7 בטבלה 10, שם העלאת הסף של מספר המתחרים גרמה להקטנת מספר הזוגות.

(ד) לטענת MBBK, אופטימיזציה של הכינויים נראית כאופטימיזציה של הפרמטרים. כדי להראות שאכן הדברים נראים כאילו היתה אופטימיזציה של הפרמטרים, ביצעו MBBK וואריאציות של פרמטרים שונים.

נתבונן למשל בפרמטר X מסויים. נסמן ב- X_0 את הערך של X בו השתמשו WRR, וב- Y_0 את תוצאת WRR עבור X_0. כדי לבדוק אם Y_0 מהווה אופטימום עבור הפונקציה $Y(X)$, צריך לדגום בנקודות X_i בסביבת X_0 ולראות האם Y_0 ערך אופטימלי לעומת הערכים $Y(X_i)$. תנאי אלמנטארי לבדיקה כזאת הוא, שהנקודות X_i יילקחו <u>משני צידי הנקודה X_0</u>. למשל, ברור שאם הפונקציה $Y(X)$ היא מונוטונית בסביבות X_0, דגימה שבה הנקודות יילקחו כך ש- $X_0<X_i$ לכל i (או ש- $X_0>X_i$ לכל i) תצביע על קילקולים בלבד או על שיפורים בלבד, למרות ש- Y_0 כלל אינו אופטימום.

לכן, על MBBK היה להראות שערכי הפרמטרים שלהם מפוזרים משני צידי הערכים של הניסוי המקורי של WRR. הם לא עשו זאת. בחלק גדול מהווריאציות דבר זה אינו ניתן להוכחה. למשל, כמעט כל השינויים שאינם בפרמטר מספרי, אלא משנים את צורת הפונקציה, אינם ניתנים לבדיקה האם הם "משני צידי" הפונקציה המקורית.

(ה) בפרק א (סעיף 6) הודגש שאחד החסרונות הבולטים באוסף הווריאציות של MBBK הן התלויות הקימות בקבוצות מסוימות של הווריאציות. ישנם מקרים בהם התלות כה חזקה, שקבוצה של וריאציות ראויה להיחשב כווריאציה יחידה.

למשל, במקרה שהפונקציה $Y(X)$ (ראה בסעיף הקודם) היא מונוטונית מצד מסוים של X_0, ידיעה של נקודת מדידה אחת מאפשרת ידיעה מראש של תוצאות שאר נקודות המדידה.

<u>לסיכום</u>: בפרק זה כללנו דוגמאות בולטות של טעויות והטעיות.

(1) מספר הווריאציות הלוקות באחד מן החסרונות המנויים בפרק זה הוא גדול מאד: למעלה מ- 100 מתוך 135 הווריאציות. רבות מהן לוקות בכמה חסרונות בעת ובעונה אחת. זאת נוסף לחסרון היסודי שהצבענו עליו בפרק א, והמשותף **לכל** הווריאציות: אוסף הווריאציות אינו סגור ומאפשר "tuning" כפי שנראה בפרק ד.

(2) דוגמאות אלו נותנות לנו מושג ברור על שיקול הדעת וההגינות ששימשו בבחירת (או יצירת) הווריאציות לצורך "מחקר הווריאציות". לאור זאת אין בסיס לבקשת MBBK לזכות באמון הקורא:

> "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)

בפרק הבא נראה זאת שוב ושוב, **ובגדול**.

פרק ג. על הצגת התוצאות ב"מחקר הווריאציות"

בפרק הקודם הראינו כי בחירת (יצירת) הווריאציות על ידי MBBK לקויה ביותר. כאן נטפל בצורת ההצגה של תוצאות הווריאציות.

בניספח C במאמרם של MBBK, ישנה סדרה מרשימה של טבלאות של תוצאות של "מחקר הווריאציות" (טבלאות 5-10 במאמרם). כבר עמדנו על כך בפרק א, כי אפילו MBBK מודים שאינם יכולים לכמת את התוצאות. בהעדר כימות, האפקט היחיד של "מחקר הווריאציות" הוא שיכנוע הקורא באמצעות "רושם", כלומר באמצעות אופן הצגת התוצאות. במובן זה יש חשיבות מיוחדת לטבלאות 5-10 שהן "חלון הראווה" של מחקר הווריאציות. אומנם, יש כמה ווריאציות בודדות נוספות המפוזרות בין שורות הטקסט של ניספח C (בחלקן כלל אינן ווריאציות בהיותן מודדות משהו אחר לחלוטין). אבל מטרתנו כאן להתמודד דווקא עם "חלון הראווה" שהם הציבו בפני הקורא, ולהראות כיצד נוצר "הרושם" המבוקש.

להצגת התוצאות השתמשו MBBK בארבע סטטיסטיקות: P2 ו- min(r1-r4) לתוצאות הרשימה הראשונה של הרבנים (L1), ו- P4 ו-min(r1-r4) לתוצאות הרשימה השניה (L2). [r_i הוא הדרוג במבחן הפרמוטציות של הסטטיסטיקה P_i.]

בפרק **א** סעיף 9 הצבענו על הליקוי היסודי הבא: הבחירה של MBBK בארבע סטטיסטיקות מסוימות כדי להציג את תוצאות הווריאציות (הקטע בשם "What measures should we compare" בעמ' 160 במאמרם), היא בחירה **א-פוסטריורית** של חלק מן המדידות. הם בחרו רביעייה אחת מתוך כמות עצומה של צרופים אפשריים, על סמך סיפורים א-פוסטריוריים. זה ליקוי עקרוני. אבל, בפרק זה נראה שהמדובר לא רק בליקוי עקרוני, אלא בסילוף של ממש.

בחלקו הראשון של הפרק נפרוש את התוצאות המלאות עבור הסטטיסטיקות השונות, ונראה שאומנם הבחירה של MBBK מסלפת באופן חמור את התמונה העולה מן הניסויים.

בחלק השני נבחן את התירוצים הא-פוסטריוריים שניתנו על ידי MBBK להצדקת השינויים שעשו. נראה שם כי לא רק שיש בהם ליקוי יסודי בהיותם א-פוסטריוריים, אלא שתירוצים אלה פשוט אינם תקפים.

1. <u>הצגת הנתונים</u>:

MBBK שואלים בכותרתם: **"What measures should we compare?"**. זו שאלה, שיתכן שהיה לה מקום אם היתה נשאלת א-פריורי. אבל לא יתכן שחוקרים שביצעו ניסויים, ישאלו באופן **א-פוסטריורי** אלו תוצאות כדאי להם להראות. התשובה ברורה: הציגו את **כל התוצאות**! כיוון ש- MBBK לא עשו זאת, שיחזרנו את כל הניסויים דלהלן, והרי אנו מציגים כאן את התוצאות.

בטבלאות 5-10 של MBBK יש רשימה של 135 וריאציות שונות. 33 הווריאציות מטבלה 5 שהוצגו על ידי MBBK למרות שאסור היה להציגן (כי הן חוזרות על וריאציה אחת: הוצאת השורש הרבועי) – נמחקו, בהתאם לאמור בפרק **ב** (חלק "ההטעיות", בסוף סעיף א). נותרו, אם כן, 102 וריאציות הנמצאות במאמרם בטבלאות 5-10. שים לב, כי 7 הווריאציות האחרונות בטבלה 10 נוגעות לפי הגדרתן רק ל-P1 (או P3) ולא ל-P2 (או P4). לכן, לגבי P2 (או P4) ישנן רק 95 וריאציות.

(א) התוצאות האמיתיות:

הבה נראה מה קורה כאשר אנו עושים את **הבחירה הטבעית** לפי התזה שלהם:

P1 ו-P2 שימשו כסטטיסטיקות היחידות להערכת הצלחת הרשימה הראשונה של הרבנים (L1) והרשימה השניה (L2). לכן, אם נעשתה אופטימיזציה, הרי נעשתה ביחס ל- P1, או ביחס ל- P2, או – מה שסביר יותר - ביחס ל- Min(P1-P2). לכן, הבחירה הטבעית היא לבדוק את התמונה ביחס לערכים אלה. והנה התוצאות:

	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

טבלה 18

"better" – אלו המקרים בהם השתפרה התוצאה עקב הווריאציה.

"equal" – אלו המקרים בהם התוצאה לא השתנתה עקב הווריאציה.

"worse" – אלו המקרים בהם התוצאה התקלקלה עקב הווריאציה.

"not worse" = "better" + "equal" – אלו המקרים בהם התוצאה לא התקלקלה עקב הווריאציה.

<u>מסקנה:</u>

התוצאות עבור Min(P1-P2) **בשני המידגמים** מצביעות על כך **שלא היתה אופטימיזציה**.

אנו מזכירים לקורא, שמחקר הווריאציות במקורו (ראה להלן פרק **ה 1(ג)**) נועד לבדוק האם נעשתה אופטימיזציה ישירה של הפרמטרים. לכן, אי הצגת תוצאות אלו על ידי MBBK תמוהה שבעתיים. אנו טוענים, שכל ההצגות שהם בחרו לתוצאות, יחד עם הסיפורים הנילווים להן – מסתירים עובדה יסודית זאת, כפי שנראה להלן.

(ב) הצגת התוצאות ב"מחקר הווריאציות":

כאן נפרוש את התוצאות המלאות עבור הסטטיסטיקות השונות, ונראה שאומנם הבחירה של MBBK מעלימה את התוצאות האמיתיות שראינו בסעיף הקודם.

(1). לפי MBBK יש מקום להציג תוצאות גם עבור הסטטיסטיקות P3 ו-P4, למרות שהן לא היו המדד למובהקות הכוללת של המידגמים. נראה מהן התוצאות עבור 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

טבלה 19

משש תוצאות אפשריות אלו בחרו MBBK את התוצאה הטובה ביותר מבחינתם: את P2, עבורו ערך "better" הוא הקטן ביותר וערך "worse" הוא הגדול ביותר.

נראה את התוצאות עבור 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

טבלה 20

משש תוצאות אפשריות אלו בחרו MBBK את... כן, נחשתם: את התוצאה הטובה ביותר מבחינתם: את P4, עבורו ערך "better" הוא הקטן ביותר וערך "worse" הוא הגדול ביותר!

<u>ניתוח התוצאות:</u>

דווקא לגבי L2, שהוא באופן מוצהר תכלית "מחקר הווריאציות" של MBBK (ראה פרק ה), **מתברר שאין שום אינדיקציה לאופטימיזציה**.

ובאשר ל-L1, אנו מבחינים במגמות סותרות: מצד אחד ב-P1, Min(P1-P4), ו-Min(P1-P2) אין אינדיקציה לאופטימיזציה. לעומת זאת, לפי המודל של MBBK יש אינדיקציה כזאת לגבי P2, P3 ו- P4. שים לב לכך, ש- P3 ו- P4 הוגדרו לראשונה הרבה לאחר הניסוי ב- L1. לכן, מוזר מאד שדווקא לגביהן ישנה אינדיקציה לאופטימיזציה, בעוד שלגבי Min(P1-P2) שהיה הקריטריון היחיד להצלחה – אין אופטימום! דיון נוסף בסתירות שנתגלו כאן, ייעשה בפרקים **ד-ה**.

(2). MBBK נתנו את המשקל העיקרי לבדיקת הווריאציות ביחס למבחן הפרמוטציות. כלומר, הם בדקו את הווריאציות לא לגבי P1 ו-P2 ששימשו כסטטיסטיקות היחידות להערכת הצלחת המידגמים המקוריים. הם עשו זאת ביחס למבחן הפרמוטציות שהוצע

שנתיים לאחר שנעשתה, לטענתם, האופטימיזציה. את הדיון בתירוצים הא-פוסטריוריים שניתנו על ידם להצדיק צעד כל כך מוזר וכל כך לא טבעי – נשאיר לחלקו השני של הפרק. כאן, נשלים את התמונה בתוצאות הווריאציות עבור מבחן הפרמוטציות. הדרוג במבחן הפרמוטציות יסומן ב- r_i.

עבור 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

טבלה 21

עבור 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

טבלה 22

עבור L2 שוב בחרו MBBK בתוצאה הטובה ביותר מבחינתם: Min(r1-r4), ועבור L1 הם העדיפו את Min(r1-r4) על r2 כדי "ליישר קו" עם הבחירה עבור L2. (הם לא יכלו לעשות כך לגבי P4 כי P4 לא הוגדר כלל עבור L1).

<u>ניתוח התוצאות:</u>

אין ספק שבמבחן הפרמוטציות התוצאות מצביעות על פחות השתפרויות מאשר בסטטיסטיקות P. אבל, התוצאות מעוררות תהיות רבות:

- שים לב, לתוצאות עבור r2 בשני המדגמים: התוצאות דומות מאד! וזאת למרות שלא רק המידגמים שונים ובנויים מביטויים אחרים, אלא שלפי התזה של MBBK צורת האופטימיזציה היתה שונה: עבור L1 היתה אופטימיזציה של הפרמטרים עצמם וגם אופטימיזציה של הנתונים וכל פרטי הניסוי. לעומת זאת עבור L2 כל הפרמטרים היו קבועים והאופטימיזציה הנטענת התרכזה בנתונים.
- ועוד יותר מפתיע הדמיון בין התוצאות עבור r4 בשני המידגמים. זאת אם נזכור שעבור L1 כלל לא הוגדרה בניסוי המקורי אותה קבוצה חלקית של כינויים, המשמשת למדידת r3 ו- r4.
- בנוסף לכך, אנו צופים במגמה מעורבת בתוצאות. הניגוד מודגש מאד עבור L2: מצד אחד r3 נותן רוב ברור להשתפרויות (ולשיטתם, הוכחה לאי אופטימיזציה), ומצד שני r4 משיג שיא בכיוון הפוך: כמעט אין השתפרויות (ולשיטתם, הוכחה לאופטימיזציה).

כל זה מעורר חשד, שלתוצאות אין שום קשר לקיומה או לאי-קיומה של אופטימיזציה! אנו נדון בכך בפרק 7.

על רקע הנתונים כולם כדאי לקרוא את דברי MBBK:

> **"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)

כמובן, אסור להתבלבל. מה שכתבו כי "the results are remarkably consistent", כוונתם היא לתוצאות **שהם רוצים להראות לנו**, ולא למשל, לתוצאות האמתיות שבסעיף **(א)** או לכלל התוצאות שראינו בחלק זה של הפרק.

2. התירוצים:

MBBK מספקים סדרת תירוצים א-פוסטריוריים כדי להצדיק את בחירת הצגת הנתונים שעשו, בעיקר בעמ׳ 160 במאמרם, תחת הכותרת:

"What measures should we compare?"

מעבר לכל הפילפולים והתירוצים, ישנן שתי שאלות יסודיות, המתבקשות מאליהן, ואשר אין כל התייחסות אליהן בדבריהם:

- מדוע הם נמנעים מלהביא את **כל** התוצאות לפני הקורא?
- מדוע, בניגוד לנדרש במחקר סטטיסטי, אינם נצמדים לאמות המידה בהן כבר בחרו בשלב קודם, אלא משנים באופן א-פוסטריורי את אמות המידה להצלחה ולכשלון?

לכן, אפילו אם היו התירוצים נכונים, השימוש בהם אינו נכון. התירוצים יכולים להצדיק לכל היותר בחירה א-פריורית של צורת ההצגה. אך אין שום הצדקה לבחירה א-פוסטריורית.

אבל לדעתנו, גם התירוצים אינם נכונים. כאן נדון בתירוצים אלה לגופם ונבחן האם הם תקפים ומה הסיבה שהביאה ליצירתם.

(א) בחירת P2 ולא P1, ובחירת P4 ולא P3:

הבה ונראה כיצד מנמקים MBBK את הבחירה הא-פוסטריורית שלהם ב- P2 עבור המידגם הראשון וב- P4 עבור המידגם השני:

הם שואלים:

> **"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?"

והם עונים:

> "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 behaviour 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."

אותו תרוץ נותנים MBBK לשלילת P_3 כמודד לווריאציות על המידגם השני:

> "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."

(1). הקורא בוודאי תמיה בקוראו תרוץ א-פוסטריורי זה, בזוכרו כי MBBK טענו **בדיוק ההיפך** בקטע קודם (בעמ׳ 155):

> **"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 P_2 and P_4 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).

אין ספק, כאשר "תופרים" תרוצים א-פוסטריוריים אין גבול לאקרובטיקה.

(2). אם ל- MBBK היה אינטרס להעדיף את P1 על P2, לא היו מתקשים להעלות טענה הפוכה: לפי הטיעון ש- P2 תלויה בכל ערכי c(w, w'), הרי זו דווקא סיבה להעדיף את P1 על P2. הסבר: הרי לפי המודל שלהם, וודאי שהאופטימיזציה נעשתה לגבי הזוגות ה"מצליחים" ולא על ה"נכשלים". לכן P2 פחות "רגיש" – כלומר, משפיעים עליו גם השינויים לגבי

ה"נכשלים". לעומת זאת, השינויים ב- P1 נגרמים בדרך כלל על ידי אוכלוסיית הזוגות המוגדרים כ"מצליחים".

אגב, להלן בסעיף **(ב)**, תמצא טענה דומה, שהעלו MBBK כאשר רצו להצדיק את העדפת P4 על P2...

מה הכריח את MBBK לבחור בטענתם ולא בטענה ההפוכה? ומה הכריח אותם בכלל **לבחור** בין השתיים?

במיוחד, נראה בסעיף **(ג)** לקמן, כי MBBK משתדלים מאד

"to capture tuning towards the objectives mentioned in the previous paragraph",

למרות שה-"objectives" האלה לקוחים מתחום הדמיון. למה הם רוצים כל כך לוותר על ההזדמנות "to capture tuning" לגבי P1 שהיה מודד מציאותי להצלחת הרשימות?

(3). קשה להבין מדוע עמלים כל כך MBBK לשלול את P1 ואת P3. אפילו נניח שאומנם הם "פחות רגישים" לווריאציות, אבל סוף סוף **אף הם צריכים להיראות כאופטימום עבור** L1 ו- L2!

אך מי שיעיין בטבלאות 19-20 לעיל יראה **למה** צריכים MBBK לשלול את P1 ואת 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

טבלה 23

פשוט מאד - כי אין כאן אופטימום!

(ב) בחירת P4 ולא P2:

MBBK ממשיכים ומסבירים:

> "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."

כאן מעלים MBBK טענה מדהימה: הם טוענים שחלק המידגם הבנוי מן כינויים הסטנדרטיים "רבי X", הוא חלק שלא נעשתה עליו אופטימיזציה, לא ישירה ולא עקיפה.

הטענה מדהימה לא בגלל שאינה נכונה, אלא משום שהיא עומדת **בסתירה גמורה** לרושם העולה ממאמרם!

<u>נסביר:</u> MBBK טורחים ומאריכים להסביר במאמרם, כמה חופש היה לנו בהכללה או באי הכללה של אישים מסויימים ב- L2, וכמה חופש היה לנו על ידי תיקון סלקטיבי של תאריכים.

הרשימה ש- MBBK מציגים עבור "מלחמה ושלום" בנויה על שלוש אופטימיזציות:

1. ביחס לכינויים. 2. ביחס להכללת אישים ברשימה. 3. ביחס לתיקון/השמטה/הוספה של תאריכים. ועל כך הם טוענים [16] שעשו בדיוק כמו WRR.

והנה, הטענה על אופטימיזציות 2. ו3. נוגעת מאד לאותו חלק של "רבי X": ישנם אפילו מקרים בהם טענה זו רלוונטית **אך ורק** לחלק זה של L2!

מה הביא אותם לטענה סותרת זו? – שתי עובדות פשוטות:

- תיקון הרכב הרשימה לפי הקריטריון שלהם [17], ותיקון/השמטה/הוספה של תאריכים לפי המומחה שלהם, היה מביא סך הכל **<u>לשיפור</u>** התוצאה המקורית, שהיא min(P1-P2), בפקטור של 3.4 [וגם אם לא נכליל את ר' דוד גנז ברשימה, לפי טענה מפוקפקת שהעלו, עדיין יש <u>שיפור</u> בפקטור של 1.8].

משמעות עובדה זו היא, ש-WRR לא עשו אופטימיזציות מסוג 2. ו3.

- אם נעמיד את קבוצת "רבי X" במבחן הוריאציות, אכן נגלה שלא היתה אופטימיזציה (את התוצאות נביא בפרק ד, סעיף 3ב).

אבל במקום להציג תוצאות אלו בפומבי, בחרו MBBK בדרך אחרת: הם בחרו להעלים מן הקורא את האמת, ולעומת זאת בחרו ליצור רושם שהיו כאן אופטימיזציות כאלה. **הסיבה לכך**: הדבר היה נחוץ להם ביותר, כי הם לא הצליחו "לתפור" רשימה ב"מלחמה ושלום" אך ורק באמצעות אופטימיזציה על הכינויים, כך שתגיע לאותה הצלחה כמו WRR. הם נאלצו להשתמש לשם כך גם באופטימיזציות 2. ו3. כדי לשפר את התוצאה בסדר גודל אחד.

מצד שני, קבוצת "רבי X" הפריעה להגיע לתוצאות הרצויות להם במבחן הוריאציות, ולהצדקת סילוקה כתבו את הקטע המצוטט לעיל.

(ג) בחירת min(r1-r4):

בשלב הבא רוצים MBBK להצדיק א-פוסטריורית את השימוש המאוחר ב- min(r1-r4). הם כותבים בהמשך דבריהם הנ"ל:

> "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_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_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.)"

יש כאן שלושה טיעונים. נפרט אותם:

<u>טענה א:</u>

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

טיעון זה לא רק שאינו רלוונטי, הוא אף מוזר: הרי, כפי שכבר הזכרנו לעיל, P1 ו-P2 שימשו כסטטיסטיקות היחידות להערכת הצלחת המידגמים המקוריים. לכן, אם נעשתה אופטימיזציה, הרי נעשתה ביחס ל- P1, או ביחס ל- P2, או – מה שסביר יותר - ביחס ל- Min(P1-P2). לכן, **א-פריורי, הבחירה הטבעית** היא לבדוק את התמונה ביחס לערכים אלה.

<u>טענה ב:</u>

> "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."

זו טענה א-פוסטריורית מובהקת של MBBK. טענה זו כוללת בעצם שלוש טענות: טענה כללית אחת ושתי טענות פרטניות:

(1). הטענה הכללית היא, שיותר נכון לבדוק את השפעת הוריאציות על ערכי r4, שהם הערכים המתקבלים על ידי מבחן הפרמוטציות, בגלל ש-r4 הוא "more reliable" מאשר P4. זו טענה שגויה. הנימוק הזה אולי רלוונטי לניסוי בו אנו מעוניינים <u>באיכות התוצאה</u>,

ולא לניסוי הוריאציות בו אנו מעוניינים לבדוק את יציבות התוצאה. ברור לגמרי, שאם אני חושד שנעשתה אופטימיזציה ביחס לסטטיסטיקה P4, כדאי לבדוק את יציבות התוצאה ביחס ל-P4, כי שם תהיה הרגישות המכסימלית לשינויים (והרי הם טענו זה עתה שהם מחפשים את הרגישות הגדולה ביותר...). לעומת זאת, לא ברור כלל כיצד אופטימיזציה על P4 תתבטא בטרנספורם (המסובך) שלו, r4, ויתכן שדווקא הפרמוטציות יהרסו חלק מן האופטימיזציה לה הם טוענים, כפי שנדגים להלן, בסעיף (3). בפרק ה נביא ראיה נסיונית לכך, שהוריאציות גרמו בהרבה מקרים לקילקול התוצאה, בגלל השפעתן על כל מיני תכונות (features) של הטרנספורם המסובך r4, וללא כל קשר לניסוי המקורי.

(2). בציטוט הנ"ל הם טענו לעדיפות r4 על P4
"in the case of experimental variations that change the number of distances".

[הערה: זו טענה שהועלתה לראשונה על ידי פרופ' גיל קלעי בנובמבר 97. מהלך הדברים היה כך: פרופ' בר-הלל השתמשה בתוצאות מבחן הפרמוטציות כדי לבחון 13 בחירות אותן הציגה בינואר 97. פרופ' אומן [9] ביקר את העובדה שהיא חישבה את "כדאיות" ו"אי כדאיות" הבחירות, לפי מבחן הפרמוטציות שהוצע כשנתיים לאחר שהבחירות נעשו, ולא בסטטיסטיקות המקוריות P1 ו- P2 אשר שימשו בניסוי המקורי, ואשר לפיהן בצעו WRR (לפי טענת בר-הלל) את ההטיות. עשרה חודשים מאוחר יותר נחלץ פרופ' קלעי לחלץ את בר-הלל ממבוכתה, ולהסביר מדוע בכל זאת צדקה כאשר נקטה במבחן הפרמוטציות המאוחר כמדד לפיו יש לחשב את ההטיה בבחירות. לפי התיעוד שבידינו [18], התקשתה בר-הלל להבין את ההסבר הרטרואקטיבי של קלעי, ולמעשה גם לא יכלה לקבלו, כי עמדתה הוצהרה בפירוש באופן אחר לגמרי.]

(א). בתת-סעיף (ב) נסביר מה בעצם כלול בטענה זו. כאן רק נבדוק האם טענה זו רלוונטית בכלל לווריאציות הנידונות במאמר הנוכחי, שהן הוריאציות שנבדקו על ידי MBBK והמפורטות בטבלאות 5-10 שבניספח C במאמרם.

מתוכן הטענה, ברור שהיא רלוונטית אך ורק לווריאציות בהם משתנה מספר הזוגות (שהוא מספר ה-distances) במידגם. הוריאציות בהן משתנה מספר הזוגות מתחלקות לשתי מחלקות:

א. וריאציות בהן קטן מספר הזוגות.
ב. וריאציות בהן גדל מספר הזוגות.

א. לגבי וריאציות בהן קטן מספר הזוגות: כפי שספרנו בפרק **ב** (חלק הטעויות, סעיף **(ג)**), ישנן בסך הכל **6** וריאציות כאלה. וכפי שהסברנו שם, זו שגיאה לקחת וריאציות בהן מקטינים את מספר הזוגות: וריאציות כאלו צפויות מראש לקלקל את התוצאה גם ב- P4 וגם ב- r4. ולכן לא היו MBBK צריכים להשתמש בוריאציות כאלה.

ב. לגבי וריאציות שבהן גדל מספר הזוגות: מתברר, ש- MBBK בדקו מעט מאד וריאציות מסוג זה:

- עבור L1: רק ב- 5 מתוך 135 הוריאציות הרשומות בטבלאות 5-10, מספר הזוגות גדל. מתברר, שבמקרים הספורים הללו, תוספת הזוגות מקלקלת כבר ב- P4 ולכן אין כאן יתרון לבדיקה ב- r4 גם לדעת קלעי [19], שטענתו באה לטפל במצבים בהם P4 משתפר ו-r4 מתקלקל!
- עבור L2: אין שום וריאציה כזאת.

לכן טענה זו כלל אינה רלוונטית לגבי הוריאציות הנידונות במאמר הנוכחי!

לסכום: אין כל הצדקה להעדיף את r4 על P4 משום טענה זו.

(ב). נסביר כאן את עצם הטענה (זו טענת קלעי [19]). הטענה מתבססת על תכונה של ביטויים מסוימים בדילוג שווה, שקראנו לה "כריזמטיות" [20]. לביטויים בעלי תכונה זו יש "יתרון" מבחינת הצלחת מפגשיהם עם אחרים.
טענת קלעי [19] היא כי האופטימיזציה (שהוא מייחס לנו) מורכבת משתיים:

I. מבחירת כינויים "כריזמטיים" (שיש להם נטייה ל"הצליח" עם כל מלה אחרת).
II. מאופטימיזציה ספציפית לגבי מפגשים עם התאריכים הנכונים.

על סמך הנחות אלו הוא מטפל במקרה הבא:
מוסיפים קבוצה חדשה של תאריכים על הקבוצה המקורית של תאריכים שהיתה כלולה במידגם, כך שכינויים "כריזמטיים" מן המידגם משתתפים עכשיו גם בזוגות הנוספים עם התאריכים החדשים. מקבלים השתפרות ב-P4. האם אפשר להסיק מכאן שלא היתה אופטימיזציה של הכינויים במידגם המקורי?
לדעת קלעי, ההשתפרות ב-P4 היא בגלל I. ולכן, למרות ההשתפרות, לא נוכל להסיק, שלא היתה אופטימיזציה.
מסקנתו היא, שבמקרים של הוספת זוגות, עלינו להשתמש במבחן הפרמוטציות המבטל את השפעת הכריזמטיות (כי כינויים "כריזמטיים" מצליחים גם עם תאריכים לא נכונים), או לפחות מחליש אותה ביותר. כלומר, לדעתו r4 עדיף במקרים אלה.
עד כאן הסבר הטענה.
כפי שהבהרנו ב(א), טענה זו לא רלוונטית לגבי הווריאציות הנידונות במאמר הנוכחי. לעומת זאת, טענה זו רלוונטית לרפליקציות (שבניספח B במאמרם) בהן משתנה, בדרך כלל, מספר הזוגות. לכן, הדיון בטענה זו ייעשה במאמרנו [10] הדן ברפליקציות (שבניספח B במאמרם). שם נביא דוגמאות נגדיות לטענה זו.

(3). MBBK מוסיפים וטוענים, ש-r4 עדיף על P4 גם במקרה של וריאציות "that tend to uniformly move distances in the same direction".

בדבריהם כאן הם מתכוונים לתכונה שקראנו לה "כריזמטיות". להזכיר, לביטויים בעלי תכונה זו יש "יתרון" מבחינת הצלחת מפגשיהם עם אחרים. והנה, בעוד שטענה **(2)** התייחסה לווריאציות המוסיפות זוגות למידגם המקורי, מתייחסת טענה **(3)** לווריאציות בהן מספר הזוגות **לא** משתנה.
טענה זו אומרת ש-r4 עדיף על P4 עבור וריאציות המגדילות את הכריזמטיות.

(א). מתברר, שדווקא לפי המודל של MBBK, טענה זו אינה נכונה **וההפך הוא הנכון!**

הבה נעיין בשני טיעונים של MBBK:
<u>טיעון א</u> (<u>טענת קלעי [19]</u>): האופטימיזציה (שהם מייחסים לנו) מורכבת משתיים:
I. מבחירת כינויים "כריזמטיים" (שיש להם נטייה ל"הצליח" עם כל מלה אחרת).
II. מאופטימיזציה ספציפית לגבי מפגשים עם התאריכים הנכונים.
<u>טיעון ב:</u> אופטימיזציה בכינויים שקולה לאופטימיזציה בפרמטרים.
טיעון א' אומר כי אכן נעשתה אופטימיזציה על ידי בחירת כינויים כריזמטיים – אופטימיזציה מטיפוס I.
מטיעון ב' נובע, כי וריאציה הגורמת ליותר כריזמטיות (ולכן לשיפור ב- P4), שקולה לבחירה של כינויים יותר כריזמטיים.
<u>מסקנה</u>: דווקא P4 הוא אינדיקטור רגיש לבדוק האם נעשתה אופטימיזציה מטיפוס I או לא. לעומת זאת, אם מבחן הפרמוטציות מבטל את הכריזמטיות, אזי r4 **כלל אינו יכול להיות אינדיקטור של אופטימיזציה מטיפוס I!** (מסקנה זו נכונה לגבי כל הסטטיסטיקות P לעומת הסטטיסטיקות r).

(ב). יתר על כן, אפילו אם רוצים משום מה לוותר על גילוי האופטימיזציה מטיפוס I, הדרך שמציעים MBBK אינה הדרך נכונה. השימוש בטרנספורם המסובך r4 לצורך זה מסלף ביותר את התוצאות (ראה בפרק ה ראיה נסיונית לכך). לעומת זאת ישנה דרך פשוטה ונכונה לנטרל את הכריזמטיות של הכינויים. בפרק ה סעיף 2**(ב)** אנו מדווחים על תוצאות ניסוי כזה. מתברר שהתוצאות שונות לגמרי מאלה המתקבלות על ידי המעבר ל- r4.

(ג). טענה חדשה (ומוטעית) זו הועלתה **לראשונה** במאמרם ב- .Stat. Sci. ייתכן שזה רק צירוף מקרים מוזר, אבל רק כאן, בשלב הנוכחי של "האבולוציה של הווריאציות", טענה זו מביאה להם תועלת. הם משתמשים בה בעמ' 169-170 כדי להצדיק מחיקת **19 השתפרויות** של P4 (מתוך 33 וריאציות) בטבלה 5. הם כותבים:

"Furthermore, in all 19 cases where P_4 dropped, the permutation rank of P_4 increased. This indicates that the observed drop in P_4 values is due to an overall tendency for $c(w, w_0)$ values to decrease when these variations are applied."

שים לב שמשפט זה אפשר **להפוך** לגבי קבוצת 19 הווריאציות הללו:

"Furthermore, in all 19 cases where the permutation rank of P_4 increased, P_4 dropped. This indicates that the observed increase in the values of the permutation rank of P_4 is due to an overall tendency for permutation ranks to increase when these variations are applied."

ואת מסקנת MBBK

"In other words, it is an example of the inadequacy of P_4 as an indirect indicator of tuning, as discussed in Section 7."

שהראינו לעיל **שאינה נכונה**, אפשר להחליף במסקנה הגיונית יותר, הנובעת מן המשפט ההפוך:

"In other words, it is an example of variations being chosen according to their destructive effect on r4, as discussed in Chapter 5 (of this paper)."

<u>טענה ג:</u>

הטענה השלישית של MBBK בעמ' 160 לעדיפות של r4 על P4 היא:

> "The permutation rank also to some extent measures P_{1_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.)"

תשובתנו:

(1). גם טענה זו שגויה מעיקרה: אופטימיזציה נוספת אינה צריכה למנוע מ- P4 המקורי להיראות כאופטימום! לעומת זאת, לא ברור כלל כיצד אופטימיזציה על P4 תתבטא בטרנספורם (המסובך) שלו r4.

(2). גם לשיטתם, תירוץ א-פוסטריורי זה נוגע רק לרשימה השניה. עבור הרשימה הראשונה לא הוכנה שום פרמוטציה ציקלית! ולכן, אם מדידת שתי אופטימיזציות מצריכה כלים **אחרים** מאשר מדידת אופטימיזציה אחת, אין שום סיבה להשתמש ב-r4, המתאים לדבריהם למדידת שתי אופטימיזציות, ברשימה הראשונה.

(3). בעצם, MBBK יכלו לטעון טיעון **הפוך**: להעדפת P4 על r4. הרי הם כותבים בפרק 8 במאמרם, כי עשינו אופטימיזציה נוספת ברשימה השניה כדי לקבל P2 קרוב מאד ל- P2 של הרשימה הראשונה. אמור מעתה, שחייבים לבדוק את הווריאציות דווקא ב- P4 ולא ב- r4 שאינו מתאים "to capture tuning towards this objective". החיסרון היחיד בטיעון ההפוך הוא, שהוא מוליך לתוצאות המראות שלא היתה אופטימיזציה...

(4). **<u>ועתה לעובדות</u>**: על איזו אופטימיזציה נוספת מדברים כאן MBBK? – לדבריהם, מדובר על כך שדאגנו לכך שלפרמוטציה הציקלית יהיה "a large value of P2 or P4" (עמ' 160).

מתברר של-P4 של הפרמוטציה הציקלית אין ערך אופטימלי אפילו מבין 31 הפרמוטציות הציקליות: הוא השלישי בגודלו. הסיכוי לכך בערך 1/10. ובאמת, אם בודקים באופן ניסיוני מה מקומו בין כלל הפרמוטציות מסוגו, פרמוטציות שאין להן חיתוך עם פרמוטצית הזהות, מתברר שדירוג שלו הוא 816 מתוך 1000. אם כן, יש לפנינו **תצפית א-פוסטריורית במאורע שההסתברות שלו היא 0.184 – והיא תצפית אחת מהרבה תצפיות א-פוסטריוריות אפשריות - זה הכל!**

המדהים הוא, שעל סמך הלא-כלום הזה בונים MBBK טיעון שלם להצדיק את העדפת r4!

לסיכום:

- כל הנימוקים שהביאו MBBK להצדקת הצעד המוזר בו **לא** נוקטים **בבחירה הטבעית**, דהיינו בסטטיסטיקות המקוריות שביחס עליהן (לטענתם) נעשתה האופטימיזציה – נועדו לאפשר הצגה מסולפת של תוצאות הווריאציות.
- כל הנימוקים הללו מופרכים.
- יש להדגיש, שגם לשיטתם, אסור היה להם לשנות מצורת ההצגה של הדו"ח המקיף הראשון – הדו"ח של מקי [5]. דהיינו, היה עליהם להציג את התוצאות עבור P1, P2, r1, r2.
- יש לזכור, שלפי המודל שלהם, גם המידות המקוריות צריכות להראות כאופטימום! אבל, כפי שהראינו לעיל, גם לפי המידות המקוריות וגם לפי כמה סטטיסטיקות נוספות אותן השמיטו MBBK, אין שום אינדיקציה לאופטימיזציה.

בפרקים הבאים נביא ראיות ניסיוניות נוספות לכך, שהווריאציות "נתפרו" כדי להציג תוצאות חריגות עבור r2 ובמיוחד עבור r4, תוצאות שמטרתן "להפליל" את WRR באופטימיזציה של הנתונים.

פרק ד. הפרכת "מחקר הווריאציות" של MBBK באופן ניסיוני

אנו נתאר כאן כמה ניסויים וחישובים שנעשו על מנת להעמיד את התזה של MBBK במבחן. אנו נדון במסקנות העולות מכל ניסוי, ולבסוף נברר האם התמונה העולה מכל הניסויים (כולל התוצאות שהובאו בפרק ג) תומכת בתזה של MBBK, או שמא היא מפריכה תזה זו.

1. <u>פרדיקציה של MBBK</u>:

במאמר קודם, של בר-הלל, בר-נתן ומקי בעיתון CHANCE [4], בו הציגו לראשונה את התזה של "מחקר הווריאציות" בלבושה הנוכחי, הם הציבו גם **פרדיקציה**:

> "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. <u>Nonetheless, our list would have fared similarly to theirs under the same checks.</u> 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).

כבר כתבנו בפרק א (סעיף 4) כי עיקר טענתם, כי "אופטימיזציה על רשימת המלים (הכינויים) צריכה להראות כמו אופטימיזציה על הפרמטרים של הניסוי" היא בגדר **סברה**. אבל, אנו עוסקים במדע ובמתמטיקה ולכן סברות צריך **להוכיח**. המפליא הוא, ש- MBBK לא ראו צורך לנסות להוכיח הנחה זו. מפליא שבעתיים כי כאן הם עצמם הצביעו על פרדיקציה הנובעת מן התיזה שלהם, אבל לא טרחו לבצע את הניסוי הבודק אם פרדיקציה זו מתאשרת.

עשינו זאת במקומם. בדקנו את השפעת הווריאציות על הרשימה ה"תפורה" שלהם ב"מלחמה ושלום". הרשימה היא זו הנתונה בטבלה 2 במאמרם (עם התאריכים שהם בחרו), ואילו הווריאציות הן המוצגות בטבלאות 5-10 (למעט 33 הווריאציות של החזקה הראשונה אותן שללנו בפרק ב, חלק ההטעיות, סעיף א). התוצאות הן:

	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

טבלה 24

באופן מוצהר, בר-נתן ומקי (BM) "תפרו" את רשימתם לפי r4 (וכך קבעו למעשה גם את Min(r1-r4) באופטימיזציה, שנעשתה בראש ובראשונה של הכינויים, אך גם מבחינת התאריכים ומבחינת הכללה או אי הכללה של אישים מסוימים במידגם. בכל זאת להשלמת התמונה, הבאנו את שאר הדירוגים. בכל הדרוגים - אין שום אינדיקציה לאופטימיזציה. גם אם נתבונן בתוצאות עבור סטטיסטיקות P (הן צריכות להעיד על אופטימיזציה עקיפה), לא נגלה סימני אופטימיזציה:

	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

טבלה 25

התמונה ברורה ביותר: אם התיזה של MBBK בעניין "מחקר הווריאציות" היא נכונה, ואם הווריאציות נבחרו (נוצרו) בצורה הוגנת – הרי לפנינו ראייה עצומה לכך שאכן טולסטוי הצפין צפנים בכוונה תחילה ב"מלחמה ושלום". וזאת, בניגוד גמור לטענת MBBK שמציאת צפנים על ידם ב"מלחמה ושלום" היא רק בגדר פרודיה!

2. <u>ניסוי נוסף: בחינת מידגם ש-MBBK מייחסים לד"ר עמנואל:</u>

בסעיף הקודם למדנו לדעת כיצד משפיעות הווריאציות של MBBK על מידגם שעבר אופטימיזציה, וראינו שהווריאציות פעלו בכיוון ההפוך מזה החזוי לפי התיזה של MBBK.

כאן נרצה לבדוק מה קורה כאשר מפעילים את הווריאציות על מידגם שעבר "טיפול" הפוך מאופטימיזציה: מידגם ש"טיפלו" בו כך, שהמובהקות שלו תורע.

(א) גם במקרה זה סיפקו לנו MBBK את המידגם. בפרק 10 במאמרם מדווחים MBBK על כמה רשימות של שמות וכינויים שהכין עבורם ד"ר עמנואל, מומחה אובייקטיבי בלתי תלוי שנשכר על ידם באופן חד צדדי. MBBK כותבים בהבלטה ובהרחבה על ניסוי שנועד "לחקות" את הניסוי של WRR. כפי שהעלתה בדיקתנו מהלך הדברים היה כך (פרטים מלאים על פרשה עגומה זו ניתן למצוא במאמרנו [8]):

(1). ד"ר עמנואל נתבקש להכין רשימת שמות וכינויים ל- 35 אישים (ביניהם 32 האישים של L2), כתחליף ל- L2, מבלי שראה את L2. רשימה זו תיקרא "רשימה ג".

(2). מרשימה זו השמיטו MBBK, ללא ידיעת עמנואל, שני אישים שהיו כלולים ב-L2. את רשימת השמות והכינויים של יתר 33 האישים, שתיקרא "רשימה גו", פירסמו MBBK בשמו של עמנואל.

(3). ד"ר עמנואל הכין גם את התאריכים הנדרשים.

(4). כך נתקבל מידגם המבוסס על השמות והכינויים של "רשימה גו". נסמן מידגם זה ב-EM3(1). מידגם זה נועד לניסוי שהוא, לדברי MBBK, בבחינת חיקוי של הניסוי של WRR.

לפי טענתם, מידגם EM3(1) נעשה באופן אובייקטיבי, ולכן לפי המודל שלהם, עקב הווריאציות צפויים להתקבל "שיפורים" באותה המידה כמו "קילקולים". ואין לטעון, שבגלל התערבותם (ראה (2) לעיל) נוצרה אופטימיזציה. אנו מפנים את הקורא למאמרנו [8] בו אנו מראים בבירור שהתערבותם נועדה לקלקל את התוצאה.

(ב) ניישם את "מחקר הווריאציות" לגבי EM3(1). הרי תוצאות "מחקר הווריאציות" עבור סטטיסטיקות-P:

	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

טבלה 26

והרי התוצאות עבור סטטיסטיקות-r:

	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

טבלה 27

שוב קיבלנו סתירה לתיזה של MBBK: הפעם רואים שרשימה שלא עברה שום אופטימיזציה (ולהיפך, קלקלו במתכוון את התוצאה), נראית ב"מחקר הווריאציות" כאילו עברה אופטימיזציה.

3. <u>ניסוי נוסף: בחינת קבוצות הכינויים מסוג "רבי X" בשני המידגמים:</u>

(א) MBBK בחרו להציג תוצאות מדידת הווריאציות לגבי Min(r1-r4) עבור L1. זה חייב אותם לבדוק את ארבע הסטטיסטיקות: r1-r4. בדיקת הסטטיסטיקות r3 ו-r4 (לפי הגדרתן) מחייבת להוריד מן המדגם קבוצה של כינויים: הכינויים הסטנדרטיים מסוג "רבי X". נסמן את המידגם החלקי המתקבל על ידי קבוצה זו ב- RABBI1. ערכי סטטיסטיקות P לגבי קבוצה זו הן:

$$P1=6.88x10^{-4},\quad P2=1.07x10^{-3}.$$

לקבוצה זו היה חלק חשוב בהצלחת L1 כולו בסטטיסטיקות P. לפי התיזה של MBBK, הצלחת L1 נבעה מאופטימיזציה ישירה על הפרמטרים של המדידה וגם מאופטימיזציה על הנתונים. לכן, יישום "מחקר הווריאציות" לגבי RABBI1 צריך, לפי התיזה שלהם, להצביע על אופטימיזציה ברורה של RABBI1, אופטימיזציה שנועדה לשפר את ערכי P1 ו/או P2 במידגם השלם, L1.

	P1	P2
better	35	50
equal	14	8
worse	53	37
not worse	49	58
total	102	95

טבלה 28

אבל התוצאות מצביעות בבירור על כך שלא היתה שום אופטימיזציה. אנו מדגישים שוב, כפי שהסברנו בפרק ג, שהבדיקה של הווריאציות צריכה להעשות בסטטיסטיקות P לפיהן נעשו הניסויים המקוריים. אנו מציגים את התוצאות עבור מבחן הפרמוטציות (סטטיסטיקות r), אך ורק להשלמת התמונה:

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

טבלה 29

גם כאן אין אינדיקציה לאופטימיזציה.
לסיכום: לפי התזה של MBBK יש כאן ראיה ברורה שלא נעשתה אופטימיזציה על הפרמטרים בניסוי הראשון, ולא על הנתונים (כינויים ותאריכים) הנוגעים לקבוצה RABBI1.

(ב) נעשה ניסוי דומה עבור L2. נסמן את המידגם החלקי המתקבל על ידי קבוצת הכינויים הסטנדרטיים מסוג "רבי X" ב- RABBI2. ערכי סטטיסטיקות P לגבי קבוצה זו הן:

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

לקבוצה זו היה חלק בהצלחת L2 כולו בסטטיסטיקות P. אומנם, MBBK אינם טוענים כי היתה כאן אופטימיזציה של הפרמטרים – שכן הם נקבעו כבר בניסוי הראשון. גם עצם הכללת קבוצת כינויים מסוג זה נקבעה בניסוי הראשון. אבל, לפי מאמריהם של MBBK [2] [16] נעשו בחלק זה לפחות שתי האופטימיזציות הבאות:

- לגבי הכללה או אי הכללה של אישים ברשימה.
- לגבי "משחק" בתאריכים.

הבה ונבחן את תוצאות הווריאציות על RABBI2:

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

טבלה 30

בולט ההבדל בין התוצאות עבור שתי הסטטיסטיקות. ההבדל נשאר גם כאשר עוברים לסטטיסטיקות r:

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

טבלה 31

על משמעות הסתירה בין התוצאות נעמוד בסעיף הבא. כאן רק נעיר, כי לפי התיזה של MBBK, צריכים להעדיף את התוצאות ב- P2 וב- r2. לכן, לשיטתם, יש כאן ראייה ברורה כי שתי האופטימיזציות הנ"ל כלל לא נעשו!

4. <u>תוצאות סותרות:</u>

ננסה לסכם את התמונה העולה עד כאן מיישום "מחקר הווריאציות" לגבי המידגמים השונים. נתייחס לתוצאות שהוצגו בסעיפים הקודמים בפרק זה, ולתוצאות המלאות עבור L1 ו- L2 שהוצגו לעיל בפרק ג חלק 1.

(א) נמיין את הסטטיסטיקות ששימשו בניסויים השונים לשני סוגים בהתאם לתוצאות שהושגו עבורן ב"מחקר הווריאציות". אם סטטיסטיקה מסוימת מצביעה על אחוז "worse" גבוה מ- 70% (סף שרירותי), נאמר שעבור אותה סטטיסטיקה יש "indication of optimization". אם המצב אינו כך, נאמר שעבור אותה סטטיסטיקה יש "No indication of optimization".

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.

טבלה 32

[<u>הערות</u>:

(א). יש לזכור שקיימת תלות חזקה בין הווריאציות. נוכח תלויות כה חזקות (שאפילו MBBK נרתעו מלכמת את התוצאות), סף של 70% או 80% הוא בעינינו סף מתון מאד. כל הנתונים הדרושים לטבלה זו מצויים במאמר, והקורא יכול לבדוק כל סף אחר, כרצונו.

(ב). כאשר מסכמים את התוצאות, כפי שנעשה בטבלה 32, יש לברר מה מעמדה של תוצאת "תיקו":

(i) נזכיר כי MBBK טוענים כי בהשערת האפס, אם **יש צופן**, צפויים 50% קילקולים מול 50% שיפורים. לפי מודל זה יש מקום להחשיב את מקרי ה"תיקו" חציים עם ה"שיפורים" וחציים עם "הקילקולים".

(ii) לפי השערת המחקר של MBBK, הטוענת שהיתה אופטימיזציה ועל כן יש לצפות דווקא לקילקולים, תוצאת "תיקו" היא נגד ההשערה יותר מאשר בעדה.

(iii) אחת משתי המטרות המוצהרות של "מחקר הווריאציות" היא מדידת היציבות של תוצאת WRR. בהיבט זה, תוצאת "תיקו" היא בבירור ראיה *נגד* יציבות.

(iv) מן האמור ב- (i) - (iii) נובע, שאם היינו צריכים לתת כימות מדוייק, יש להניח שהמשקל של תיקו אינו כמו המשקל של שיפור, והוא נמצא בין 0.5 ל- 1.

(v) במצב הקיים, כאשר הסיכום נועד בעיקרו לתת אומדן גס, סיכמנו בטבלה 32 את התיקו עם השיפורים (משקל של 1 ל"תיקו"). ולהשוואה, ניתנת כאן טבלה 32א, כאשר הנתונים בה מחושבים לפי משקל של 0.5: חישבנו את מקרי ה"תיקו" חציים עם ה"שיפורים" וחציים עם "הקילקולים":

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.

טבלה 32א

מתברר שהשינוי המשמעותי היחיד הוא עבור EM3(1). סוף ההערות].

הדבר היחיד הברור מטבלה 32 הוא שהנתונים המוצגים בה סותרים זה את זה, ואינם מתיישבים בשום אופן עם "מחקר הווריאציות".

(1). התוצאות שנתקבלו עבור BM Sample, מידגם שבו נעשתה אופטימיזציה באופן מוצהר, אינן מראות שום סימן של אופטימיזציה. "מחקר הווריאציות" קרס כאן באופן בולט.

(2). התוצאות שנתקבלו עבור EM3(1), מראות גם אינדיקציה לאופטימיזציה, למרות שמידגם זה לטענת MBBK לא עבר אופטימיזציה, ולאמיתו של דבר אפילו עבר "טיפול" הפוך: להרעת המובהקות.

(3). עצם הסתירות בין התוצאות עבור L1 וכן עבור L2, מצביעות על אותה מסקנה. והדברים רק מחריפים אם מתייחסים לטענות הצדדים:

(א). אם, כפי טענת MBBK, נעשתה אופטימיזציה במידגמים אלה, אזי:

- מדוע דווקא בסטטיסטיקה Min(P1-P2), שביחס אליה היתה צריכה להיעשות כביכול האופטימיזציה - אין אינדיקציה לאופטימיזציה, ולעומת זאת, לגבי P4 שהוגדרה לראשונה הרבה **לאחר** הניסוי ב- L1, ישנה אינדיקציה לאופטימיזציה ב- L1?
- מדוע רוב הסטטיסטיקות אינן מראות אופטימיזציה לגבי L2, ובייחוד אלה שביחס אליהם נעשתה, כביכול, האופטימיזציה. ודווקא הסטטיסטיקה r4 (ושלוש סטטיסטיקות התלויות בה תלות חזקה - r2, Min(r1-r2), Min(r1-r4)) היא זאת שמצביעה על אופטימיזציה למרות שלא ביחס אליה היתה אמורה להיעשות האופטימיזציה?
- ומדוע לשיטתם אין אינדיקציה לאופטימיזציה במידגם RABBI1? ומדוע לגבי RABBI2, המהווה חלק מ-L2, אין אינדיקציה לאופטימיזציה בסטטיסטיקה המצביעה על אופטימיזציה ב- L2 ?

זאת, בנוסף לתמיהות נוספות שכבר הסבנו אליהן את תשומת הלב בפרק ג.

(ב). אם, כטענתנו, שלא היתה אופטימיזציה ב- L1 ו- L2, איך יתכן שכמה סטטיסטיקות מראות כאילו היתה אופטימיזציה?

(ב) כדי לנתח מדוע "מחקר הווריאציות" הניב תוצאות מוזרות אלה, נזכר קודם כל בשתי הטענות העיקריות עליהן הוא נשען:

(1). אופטימיזציה על הנתונים נראית כאופטימיזציה על הפרמטרים.

(2). דגימת הווריאציות היתה באופן נכון והוגן.

מנתוני הטבלה נובע כי לא יתכן שגם (1) וגם (2) נכונים. לכן, ברור כי נותרו שלוש אפשרויות בלבד:

I. (1) אינו נכון.

II. (2) אינו נכון.

III. גם (1) וגם (2) אינם נכונים.

<u>בדיקת השערה I:</u>
נבדוק את ההשערה שהגורם <u>היחיד</u> לתוצאות הנ"ל הוא אי-תקפות (1). לפי השערה זו מובנת התמונה המתקבלת עבור BM Sample ו- EM3(1), ואף אפשר ש- MBBK יהיו מעוניינים להסביר לפי השערה זו גם את התמונה המתקבלת עבור RABBI1 ו- RABBI2.

אבל אין בהשערה זו כדי להסביר מדוע דווקא לגבי L1 ו- L2 יש סתירה כה בולטת בתוצאות, ומדוע ישנו **דמיון רב** בסתירות אלה:

(א). בשני המידגמים, דווקא בסטטיסטיקה Min(P1-P2), שביחס אליה היתה צריכה להיעשות כביכול האופטימיזציה - אין אינדיקציה לאופטימיזציה, בעוד שבשניהם, האינדיקציה העיקרית לאופטימיזציה באה מן הסטטיסטיקה r4 (ושלוש סטטיסטיקות התלויות בה תלות חזקה - r2, Min(r1-r2), Min(r1-r4)).

(ב). באופן מוזר, התוצאות עבור r2 בשני המדגמים דומות מאד, וזאת למרות שהמדגמים שונים ובנויים מביטויים אחרים, ולמרות ש(1) אינו נכון! המוזרות בולטת יותר אם נזכור שלפי התזה של MBBK צורת האופטימיזציה בשני המידגמים היתה שונה: עבור L1 היתה אופטימיזציה של הפרמטרים עצמם וגם אופטימיזציה של הנתונים וכל פרטי הניסוי. לעומת זאת עבור L2 כל הפרמטרים היו קבועים והאופטימיזציה הנטענת התרכזה בשמות והכינויים.

(ג). עוד יותר מפתיע הדמיון בין התוצאות עבור r4 בשני המידגמים. זאת אם נזכור שעבור L1 כלל לא הוגדרה בניסוי המקורי אותה קבוצה חלקית של כינויים, המשמשת למדידת r3 ו- r4.

<u>מסקנה:</u> אי אפשר להסביר את התוצאות בטבלה לפי השערה זו.

<u>בדיקת השערה II:</u>
נבדוק את ההשערה שהגורם <u>היחיד</u> לתוצאות הנ"ל הוא אי-תקפות (2). כלומר, עלינו לבדוק האם התמונה העולה מטבלה 32 היא אך ורק תולדה של דגימה לא תקינה של הווריאציות. שתי סיבות אפשריות לדגימה לא תקינה:

א. טעות: בחירה לא-מכוונת של וריאציות שאינן רלוונטיות מסיבות שונות, היווצרות תלויות בין הווריאציות וכו'.

ב. "תפירה" (tuning): בחירה מכוונת של וריאציות כדי להגיע לתוצאה מבוקשת.

<u>בחינת אפשרות א:</u>
אם כל הכשל של "מחקר הווריאציות" נעוץ רק בבחירה לא-מכוונת של וריאציות "לא נכונות", קשה להסביר:

- מדוע בחירה זו "הרעה" דווקא למידגמים L1 ו- L2, שסומנו כמטרה על ידי MBBK (ואגב כך למידגם EM3(1) המכיל הרבה זוגות "כינוי-תאריך" של L2), אבל "היטיבה" עם האחרים, ובמיוחד עם BM Sample.
- את כל התמיהות שהעלנו כנגד השערה I.

לכן, אין באפשרות זו להסביר את התוצאות.

<u>בחינת אפשרות ב:</u>
לפי אפשרות זו, התוצאות בטבלה 32 הן תולדה של "תפירה" – בחירה מכוונת של וריאציות כדי להשיג תוצאה רצויה. התוצאה הרצויה היא זו המודגשת על ידי MBBK, שהסטטיסטיקות שהם בחרו להציג עבור L1 ו- L2, נותנות אינדיקציה לאופטימיזציה. אפשרות זו מסבירה:

- את התאימות הרבה בין הסטטיסטיקות שנבחרו על ידי MBBK לבין היותן אינדיקציה לאופטימיזציה עבור כל אחד משני המידגמים הנ"ל (כפי שהראינו בחלקו הראשון של פרק ג).
- את התמיהות (א)-(ג) שהעלנו כנגד השערה I (שהן גם התמיהות שהעלנו נגד "אפשרות א" בהשערה הנוכחית):

(א). MBBK כיוונו את מאמצי ה"תפירה" של הוריאציות כדי להרוס את תוצאת WRR, דהיינו Min(r1-r4), וזו הסיבה שסטטיסטיקה כמו Min(P1-P2) לא נפגעה כל כך. כפי שנראה בפרק הבא, הוריאציות גורמות לקילקול התוצאה על ידי פגיעה בעיקר במבחן הפרמוטציות, ולכן פגעו הרבה פחות בסטטיסטיקות-P. נוסף על כך, הוריאציות כוונו לפגוע יותר בסטטיסטיקות המבוססות על P2 (P4) ולא בסטטיסטיקות המבוססות על P1 (P3).

(ב)-(ג). הדמיון בין התוצאות של r2 ו-r4 עבור שני המידגמים L1 ו- L2, משקף את הציפיות הנאיביות של ה"תופרים".

- את התוצאות עבור EM3(1), מידגם המכיל הרבה זוגות "כינוי-תאריך" של L2, ודווקא מאותה הקבוצה ש- P4 ו- r4 מוגדרות לגביה (הרחבה על כך בפרק ה, 2 **(ג)**).

אפשרות זו נתמכת על ידי הראיות שהבאנו בפרקים ב-ג ל"תפירה" של וריאציות ולבחירה מכוונת שלהן. אפשרות זו גם תואמת את טענתנו כי לא נעשתה אופטימיזציה על L1, L2, RABBI1, RABBI2, ולכן:

- עבור סטטיסטיקות אחרות מאלו שהציגו MBBK לגבי L1 ו- L2, אין אינדיקציה לאופטימיזציה (חוץ ממקרה בו יש תלות חזקה בסטטיסטיקות שהם בחרו להציג).
- לגבי RABBI1 ו- RABBI2 אין אינדיקציה לאופטימיזציה בסטטיסטיקות שנבחרו להצגה על ידי MBBK והרלוונטיות לגבי מידגמים אלה: P2, r2.

נותר עדיין לדון רק בתוצאות עבור BM Sample: מדוע כאן אין אינדיקציה לאופטימיזציה. הסבר אפשרי הוא, שהבחירה המכוונת של וריאציות לצורך "הוכחת" אופטימיזציה ב- L1 ו- L2, יצרה אוסף לא תקין של וריאציות (למשל, על ידי יצירת תלות בין הוריאציות), שהוביל לתוצאה מעוותת עבור BM Sample.

<u>מסקנה</u>: יש אפשרות להסביר את התוצאות בטבלה 32 כתולדה של "תפירה" (tuning): בחירה מכוונת של וריאציות כדי להגיע לתוצאה מבוקשת.

<u>בדיקת השערה III:</u>

נבדוק את ההשערה שהתמונה המתקבלת מטבלה 32 נוצרה מצרוף אי-תקפותם של (1) ושל (2) גם יחד. כלומר נבדוק את האפשרות שהתוצאות הנ"ל הן תולדה של כשלון השערת MBBK שאופטימיזציה על הנתונים נראית כאופטימיזציה על הפרמטרים, יחד עם דגימה לא תקינה של וריאציות. לפי הדיון בשתי ההשערות הקודמות, נשאר לבדוק רק את צרוף שתי הסיבות הבאות:

א. אופטימיזציה על הנתונים אינה נראית כאופטימיזציה על הפרמטרים.

ב. "תפירה" (tuning).

ראינו לעיל שדי בסיבה ב. ("תפירה") לבדה כדי להסביר את התוצאות בטבלה 32. ולכן, גם הצרוף של א+ב יש בו כדי להסביר זאת.

<u>מסקנה</u>: יש אפשרות להסביר את התוצאות בטבלה 32 גם לפי השערה III.

<u>לסיכום</u>:

ההסבר לנתונים בטבלה 32 הוא, שהם מהווים תולדה (ישירה ועקיפה) של "תפירה" (tuning) ובחירה מכוונת של וריאציות לצורך "הוכחת" אופטימיזציה ב- L1 ו- L2.

לעומת זאת, אי אפשר להכריע מנתונים אלה את תקפות השערת MBBK, כי אופטימיזציה על הנתונים נראית כאופטימיזציה על הפרמטרים.

בפרק הבא נחקור ביתר פירוט כיצד התקבלו התוצאות של MBBK עבור L1 ו- L2. נעשה זאת על ידי הצעת מודל אלטרנטיבי לתיזה שלהם, והעמדתו במבחנים ניסיוניים נוספים.

פרק ה. מודל אלטרנטיבי: האבולוציה של "מחקר הווריאציות"

בפרק הקודם ראינו כי ישנן סתירות חמורות בין תוצאות הווריאציות בניסויים השונים, ניתחנו את הגורמים לכך, והסקנו כי היתה "תפירה" (tuning) ובחירה מכוונת של ווריאציות לצורך "הוכחת" אופטימיזציה ב- L1 ו- L2.

בפרק זה אנו רוצים לצעוד צעד אחד קדימה, ולהציע מודל להסברת התוצאות הניסיוניות. אנו רוצים לבחון את האפשרות שתהליך בחירת (יצירת) הווריאציות היה כזה, שנועד להעמיד במבחן בעיקר את התוצאה היחידה שפורסמה על ידי WRR: את התוצאה של מבחן הפרמוטציות (Min(r1-r4)) עבור הרשימה השניה (L2). זה מודל אלטרנטיבי למודל של MBBK.

"מחקר הווריאציות" עבר תהליך אבולוציוני ממושך ומפותל. האבולוציה היתה כפולה: הן בהוספה (או השמטה) בלתי מבוקרת של ווריאציות, והן בבחירה מה לפרסם ומה להסתיר.

בחלקו הראשון של הפרק נראה כיצד השתנתה התמונה בהדרגה לטובת מטרתם העיקרית של MBBK: להוכיח שהיתה אופטימיזציה מכוונת בבחירת הכינויים לניסוי WRR על הרשימה השניה.

בחלקו האחר של הפרק נביא תוצאות של ניסויים נוספים המצביעים על כך שהתוצאה של MBBK היא מלאכותית: שאין קשר בין התזה שלהם לגבי האופטימיזציה, לבין התוצאה שהם מציגים עבור L2.

כדי להציג תמונה שלמה ככל האפשר מבלי להטריח את הקורא לדפדף בפרקים הקודמים במאמר, חזרנו והצגנו כאן גם כמה עניינים ונתונים המפוזרים במקומות אחדים במאמר.

1. <u>האבולוציה של נתוני MBBK</u>:

במבוא למאמרם מתארים MBBK בקצרה את הניסוי של WRR על הרשימה השניה של הרבנים (L2), ומצטטים את תוצאת הניסוי. משם ואילך, בהזכירם בצורה סתמית את המלים "ניסוי" או "תוצאה" (של WRR) הם מתכוונים לניסוי זה ולתוצאה זו. ניסוי זה הוא הנושא של מאמרם:

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

בהסבירם את מטרת "מחקר הווריאציות", הן במבוא (עמ' 151-152) והן בראש הפרק השביעי במאמרם, מתייחסים MBBK אך ורק לניסוי על L2.

עיקר החקירה שלהם היא, אם כן, נגד L2. ובאמת, כבר היו צריכים להביא סיבה כדי לכלול גם את חקירת L1. בסוף הפרק השלישי במאמרם הם צריכים להתאמץ ולהסביר:

> "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)

לכן, נתמקד כאן בבחינת האבולוציה של נתוני MBBK בנוגע לתוצאה של L2.

(א) <u>התוצאות האמיתיות</u>:

הבה נראה מה קורה כאשר אנו עושים את **הבחירה הטבעית** לפי התזה שלהם:

- את השפעת הווריאציות אנו בודקים על L2.
- P1 ו-P2 שימשו כסטטיסטיקות היחידות להערכת הצלחת המידגמים המקוריים. לכן, אם נעשתה אופטימיזציה, הרי נעשתה ביחס ל- P1, או ביחס ל- P2, או – מה שסביר יותר - ביחס ל- Min(P1-P2). לכן, הבחירה הטבעית היא לבדוק את התמונה ביחס לערכים אלה.

והרי התוצאות עבור 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

טבלה 33

אין כאן שום ראייה עקיפה לאופטימיזציה! אדרבא, להיפך: אם התזה של MBBK בעניין "מחקר הווריאציות" נכונה, הרי **יש כאן ראיה ברורה שלא היתה אופטימיזציה!**

אנו טוענים, שכל ההצגות שהם בחרו לתוצאות, יחד עם הסיפורים הנילווים להן – מסתירים עובדה יסודית זאת, כפי שנראה להלן.

(ב) <u>מוטציה</u>:

התוצאות של הבדיקה הטבעית במסגרת "מחקר הווריאציות" עומדות **בניגוד משווע** לדיווח של MBBK בכמה פירסומים, לפיו התוצאה מתקלקלת כמעט תמיד תחת הווריאציות. למשל **ב-** 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)

כדי להבין את מה גרם לדיווח השונה, נביא עוד קטע [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)

סוף הציטוט מסגיר את מקור ההבדל בדיווח. הם בדקו את הווריאציות לא לגבי P1 ו-P2 ששימשו כסטטיסטיקות היחידות להערכת הצלחת המידגמים המקוריים. הם עשו זאת ביחס למבחן הפרמוטציות שהוצע **שנתיים לאחר** שנעשתה, לטענתם, האופטימיזציה. את הדיון בתירוצים הא-פוסטריוריים שניתנו על ידם להצדיק צעד כל כך מוזר וכל כך לא טבעי – ערכנו בחלקו השני של פרק ג. כאן רק נסכם את התוצאות שהם בחרו להציג עבור L2:

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

טבלה 34

את עיקר הדגש שמו MBBK על הטור הימני, והם כותבים:

> **"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. <u>This trend is most extreme for the permutation test in the second list, the only success measure presented in WRR94</u>." (Pg. 169, emphasis mine)

יש לציין, כי לפי "הטעות" שעשו בהביאם בחשבון 33 וריאציות נוספות, שכולן חזרה על אותה וריאציה (של הוצאת שורש ריבועי, ראה פרק **ב 1(א)**) – התוצאה עוד יותר חדה מבחינתם: רק 4 השתפרויות מתוך 135 וריאציות! ערך זה תואם את הציפיות הנאיביות שלהם כפי שהן מובעות בציטטות הקודמות מ- CHANCE, או במאמרם הנוכחי (בפרט בעמ' 169).

(תזכורת: מה שכתבו כי "the results are remarkably consistent", כוונתם היא לתוצאות שהם רוצים להראות לנו, ולא למשל, לתוצאות האמיתיות שבסעיף הקודם).

לאור ההבדל העצום בין טבלה 33 לטבלה 34, נשאלת השאלה: מהי המוטציה שגרמה לתמונה המאוזנת שראינו בסעיף הקודם, להפוך לתמונה כל כך חד-צדדית באופן קיצוני?
כדי לענות על שאלה זאת, נתחקה אחר התהליך האבולוציוני עצמו.

(ג) **<u>אבולוציה בפעולה</u>:**
ראשית נשרטט בקצרה את השלבים הידועים לנו ב"מחקר הווריאציות".

1 העבודה של בר-הלל, ינואר 97. בדקה את כדאיותן של 13 בחירות שנעשו לדעתה בניסויים על שתי רשימות הרבנים. הבדיקה נעשתה על ידי r1 ו- r2 ועל ידי עוד שתי סטטיסטיקות ש-WRR לא השתמשו בהן מעולם. התוצאה עליה הכריזה [21] היא 0 מול 13 לרעת WRR. את ביקורתו הנוקבת של פרופ' אומן על עבודה זו תוכל לקרוא כאן [9]. בי"ז טבת התשנ"ט, בעת דיון בעקבות הרצאתנו ב"מרכז לחקר הרציונאליות" באוניברסיטה העברית בירושלים, הודיעה פרופ' בר-הלל כי זרקה את עבודתה ל- "waste basket".

2 הדו"ח הראשון של מקי, פברואר 97. מקי בדק רק "robustness" של L2 עבור ארבעה מקרים, באמצעות r1 ו- r2. הוא בדק זאת גם בשלוש סטטיסטיקות נוספות ש-WRR לא השתמשו בהן מעולם.

3 הדו"ח השני של מקי, אפריל 97. זו בעצם העבודה הגדולה הראשונה בנושא. כאן מקי שואל "Where is the low probability", ומנסה להראות באמצעות בחינת ווריאציות, שיש חופש תימרון גדול בפרמטרים של הניסוי. כאן הרציונאל של השימוש בווריאציות הוא: אם יתברר שיש הרבה ווריאציות הנותנות תוצאות גרועות יותר, המסקנה תהיה שהתוצאה בניסוי המקורי (0.00002) התקבלה בגלל הבחירות שנעשו בניסוי. הוא בדק 20 קבוצות של ווריאציות, ברובן יש יותר מנקודת מדידה אחת. הוא מציג את התוצאות עבור P1 ו- P2, ו- r1 ו- r2 לגבי שתי רשימות הרבנים.

4 המאמר של BBM (בר-הלל, בר-נתן ומקי) ב- CHANCE, אביב 98. כאן ישנה הצגה סלקטיבית של ווריאציות להדגמה, חלקן הוצגו כאן לראשונה. הכל נמדד לפי r2 בלבד, ללא שום איזכור שנעשתה מדידה לפי סטטיסטיקות אחרות וללא שום הנמקה. הדגש הוא על L2, ועבור רוב הווריאציות התוצאות ניתנות גם עבור L1.

5 המאמר הנוכחי של MBBK ב- *Statistical Science*, מאי 99. כאן הוצגו לראשונה ווריאציות רבות. אומנם לא רבות כל כך ביחס לכמות של <u>מאות</u> ווריאציות שנבדקו על ידי MBBK לפי עדותם: כשנה קודם לכן כבר כתבו [6] שמספר הווריאציות "…<u>by now</u> number in the <u>hundreds</u>". הפעם בחרו MBBK להציג את התוצאות בסטטיסטיקות הבאות: עבור L1 ב-P2 ו- min(r1-r4) ועבור L2 ב-P4 ו- min(r1-r4).

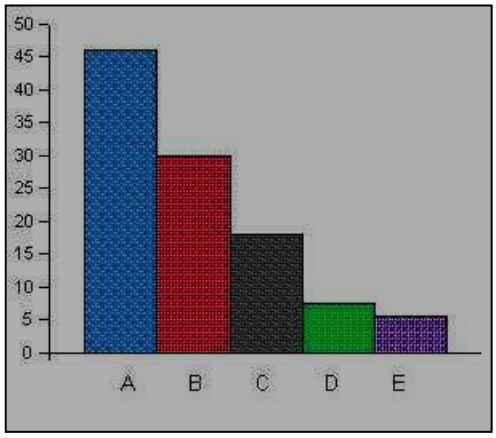
היה כאן תהליך של התפתחות גם בווריאציות עצמן וגם בהצגת תוצאות המדידה. כדי להבין כיצד הגיעו MBBK לטבלה 34, נעקוב מה קרה בשלבים הבאים לתוצאות הווריאציות עבור L2 **לפי מבחן הפרמוטציות:**

<u>נסמן</u>:
A=בסיס להשוואה: התוצאות האמיתיות עבור min(P1-P2) (ראה טבלה 33).
B=צורת ההצגה של שלב 3 (r1 ו- r2). הווריאציות הן אותן הווריאציות עבורן יש תוצאות בדו"ח של מקי, ואשר נמצאות גם בטבלאות 5-10 במאמרם הנוכחי.
C=צורת ההצגה של שלב 3 (r1 ו- r2). הווריאציות הן כל הווריאציות הנמצאות בטבלאות 5-10 במאמרם הנוכחי.
D=צורת ההצגה של שלב 4 (r2). הווריאציות הן כל הווריאציות הנמצאות בטבלאות 5-10 במאמרם הנוכחי.
E=צורת ההצגה של שלב 5 [min(r1-r4)]. הווריאציות הן כל הווריאציות הנמצאות בטבלאות 5-10 במאמרם הנוכחי.

נציג את אחוז הוריאציות בהן השתפרה התוצאה במבחן הפרמוטציות בשלבים השונים (התוצאות הן לפי מה ש-MBBK רצו להראות לנו: כלומר, לפי השערת האפס המוטעית של MBBK. במקרה זה אנו מחלקים בשווה את מקרי ה"תיקו" בין "השיפורים" ו"הקלקולים"):

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

טבלה 35



הגרף של ירידת אחוז השיפורים עם חלוף השנים.

לאחר שעקבנו אחרי התהליך האבולוציוני עצמו, נותר לברר כיצד נוצרה המוטציה שראינו בטבלה 34. זאת נעשה בסעיפים הבאים.

2. היכן קבור הכלב?

כאן אנו מתחילים לברר, בדרך האלימינציה, היכן בדיוק נמצאים אותם הנתונים השייכים ל-L2, אשר לפי התזה של MBBK נעשתה לגביהם אופטימיזציה.

(א) כאשר MBBK מנתחים במאמרם את אפשרויות "התפירה" של נתוני הרשימה השניה (L2), הם מונים שלושה רכיבים המאפשרים את "התפירה" ההיפותטית:

א. **החופש בבחירת האישים** ("The choice of rabbis" - **בעמ' 155 במאמרם**).

ב. **החופש בבחירת התאריכים** ("The choice of dates" - **בעמ' 551 במאמרם**).

ג. **החופש בבחירת הכינויים** ("The choice of appellations" - **בעמ' 561 במאמרם**).

(1). מבין שלושה רכיבים אלה רואים MBBK כרכיב העיקרי את <u>רכיב ג</u>: "**החופש בבחירת הכינויים**". הם מדגישים זאת לאורך מאמרם. למשל, **בתחילת פרק 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)

הם טוענים כאן, כי די <u>ברכיב ג</u> **כדי להסביר את הצלחת** L2.

לעומת זאת, אפשר לדעת בנקל כי לא היה שום ניצול של רכיבים א' וב' לצורך שיפור התוצאה שלנו: קביעת הרכב האישים ברשימה השניה לפי הקריטריון שלהם להכללת אישים, יחד עם השימוש בתאריכים של המומחה שלהם, נותנים רשימה שהצלחתה **גדולה יותר**.

נפרט: קביעת הרכב האישים ברשימה השניה לפי הקריטריון של MBBK להכללת אישים [16] [17], והשימוש **בתאריכים לפי המומחה שלהם, נותנים רשימה שהצלחתה גדולה יותר** [לפי המודד שהיה באותה העת: min(P1-P2)]. **<u>השיפור</u>** בתוצאה הוא **בפקטור של 3.4.**

[הערה: גם אם לא נכליל את ר' דוד גנז ברשימה, לפי טענה מפוקפקת שהם העלו, עדיין יש **<u>שיפור</u>**

בפקטור של 1.8, וגם אם נפעל לפי הקריטריון שהמציאו MBBK זה עתה להכללת אישים ([8] פרק א, 2(ב)(1)) גם אז מקבלים **שיפור** בתוצאה, אומנם בפקטור צנוע יותר].
ולכן, MBBK עצמם, בהצבעם על הקריטריון להכללת אישים ברשימה, ובהצעם תיקוני תאריכים, איפשרו להוכיח כי ביצירת הרשימה השניה **הבחירות שלנו היו לרעתנו**, מה שמוכיח כי הן נעשו בתום לב.

(2). נותרנו, אם כן, עם רכיב ג, אשר לפיו החופש בבחירת השמות והכינויים ב- L2 הוא שאיפשר את האופטימיזציה. האם, לפי טענתם, היה חופש כזה לגבי כל הכינויים ב- L2?

מתברר, כי גם לפי טענתם (ראה פרק **ג** חלק 2 (ב)) לא נעשתה אופטימיזציה ביחס לקבוצת הכינויים הסטנדרטיים מסוג "רבי X". נסמן את קבוצת **שאר** הכינויים של L2 ב- L'2. אם כן, לפי טענת MBBK נעשתה אופטימיזציה דווקא על כינויי L'2, והראיה שהם מביאים היא התוצאה החריגה min(r1-r4). התוצאה של min(r1-r4) ב"מחקר הווריאציות" נובעת כולה מן התוצאה עבור הסטטיסטיקה r4, וסטטיסטיקה זו לפי הגדרתה מודדת רק את קבוצת הכינויים L'2 (הסטטיסטיקות המתייחסות ל- L'2 הן אך ורק: P3, P4, r3, r4).
לכן עלינו לחקור: איפה בדיוק ב- L'2 נמצאת "האופטימיזציה"?

(ב) ראשית, נעתיק את התוצאות עבור L'2 מטבלאות 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

טבלה 36

מן הטבלה ברור כי רק r4 מצביעה על אופטימום. MBBK העלו טענות שונות ומשונות להעדיף את r4 על P4. בפרק **ג** (חלק ה"תרוצים") כבר טיפלנו בכל התירוצים הא-פוסטריוריים שהביאו MBBK לעניין זה. ראה שם, כי אחת מטענותיהם היתה, שהשיפורים ב- P4 נובעים מ"נטייה" מסוימת, שאנו קראנו לה ה"כריזמטיות" של הכינויים. ולכן לדעתם, r4 היא הסטטיסטיקה הנכונה, משום שהיא מקזזת את ה"כריזמטיות". ראה שם כיצד פרכנו את טענתם. במסגרת החקירה הנוכחית, אנו רוצים לבדוק האם באמת ההבדל הגדול בתוצאות בין r4 ל- P4, נובע מסיבה זו.

(1) ישנה דרך פשוטה לעשות זאת: לחשב את ערכי המפגשים של התאריכים עם הכינויים, כאשר הכינויים נלקחים אך ורק בדילוג השווה. כך מתבטל כל האפקט של הכריזמטיות של הכינויים [20]. חזרנו ובצענו את הווריאציות כאשר המפגשים מחושבים בדרך זו, והסטטיסטיקה היא P4. להשוואה חישבנו גם את ערכי המפגשים כאשר הפעם התאריכים הם הנלקחים אך ורק בדילוג שווה:

	תאריכים רק בדילוג שווה	חישוב רגיל	שמות רק בדילוג שווה
better	26	31	29
equal	7	7	7
worse	60	57	57
not worse	33	38	36
total	93	95	93

טבלה 37

[עבור הטורים הימני והשמאלי יש רק 93 וריאציות, כי שתי וריאציות אינן אפשריות עבור "שמות רק בדילוג שווה" ו"תאריכים רק בדילוג שווה"].

אנו רואים שהתוצאות דומות, ואין הבדל מהותי ביניהן. לעומת זאת התוצאה לגבי r4 שונה בתכלית. זה מוכיח כי התוצאה ב- r4 נובעת מגורמים אחרים, אותם ננסה לברר בהמשך. בזה הפרכנו את טענת MBBK באופן ניסיוני. גם בשאר טענותיהם לא הצליחו MBBK להצביע אפילו על סיבה אחת נכונה להעדיף את r4 על P4 (ראה בפרק **ג**, חלק ה"תרוצים").

(2) עצם ההבדל הגדול בין תוצאות הוריאציות עבור P4 לאלו עבור r4, נראה חריג ביותר. הטבלה הבאה תמחיש זו. נסמן ב- Imp(P4) את מס' השיפורים עבור P4, וב- Imp(r4) את מס' השיפורים עבור r4. נגדיר: Q=Imp(P4)/Imp(r4). אזי נקבל עבור המידגמים השונים (ראה פרק ד):

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

טבלה 38

(לגבי RABBI1 ו- RABBI2 נלקחו P2 ו- r2, כי P4 ו- r4 לא מוגדרים עבורם). לדעתנו, התוצאה החריגה מאד עבור L2, היא כשלעצמה תוצאה של ה"tuning" בוריאציות, כאשר החריגה עבור L1, שהיא קטנה ממנה היא אפקט לוואי. ועניין זה שווה חקירה נוספת.

<u>לסיכום</u>: לדעתנו הדרך הנכונה היא להשתמש לצורך "מחקר הוריאציות" בסטטיסטיקות-P, ולכן לגבי L'2 הבדיקה הנכונה היא עם P3 ו-P4. רק כדי להמשיך ולברר כיצד "נתפרו" הוריאציות נראה גם את התוצאות בסטטיסטיקות-r (r3 ו- r4).

(ג) במאמרם, בפרק 10, מדווחים MBBK על רשימות שמות וכינויים שהוכנו עבורם על ידי ד"ר שמחה עמנואל. פירסמנו מאמר מיוחד [8] המטפל ברשימות אלה. אחת מרשימות הכינויים הללו, שקראנו לה "רשימה ג", נועדה לחקות את L2. כאן נערוך ניסוי בעזרת רשימה זו.

לצורך הניסוי, ניקח מתוך "רשימה ג" את הכינויים של 32 האישים של L2, אשר הם בני 5-8 אותיות (כפי שנעשה בניסוי המקורי). נסמן קבוצה זו ב- EM3. נכין שתי קבוצות כינויים:

- קבוצה A, היא החיתוך של EM3 עם L'2.
- קבוצה B = L'2-A.

הקבוצה EM3 לא עברה אופטימיזציה ולכן לפי התיזה של MBBK אין לצפות לגלות סימני אופטימיזציה בהפעלת הוריאציות.

קבוצה A דומה מאוד ל- EM3: היא מכילה אותם שמות וכינויים שבחר ד"ר עמנואל, למעט כמה כינויים בודדים [ישנם רק 6 כינויים כאלה (או צורת כתיב אחרת של אותו כינוי), ומתוכם אחד שאינו מופיע כ- ELS בבראשית, ולכן אינו רלוונטי לדיון הנוכחי]. לכן, סביר שבמבחן הוריאציות תתנהג בצורה דומה ל- EM3.

לעומת זאת B מכילה את שאר הכינויים של L'2: אותם הכינויים בהם בחר הבלין ולא בחר עמנואל. לכן **אם** היתה אופטימיזציה על הכינויים, היא היתה על הכינויים ב- B.

נבדוק את הצלחת הקבוצות EM3, A ו- B לפי הסטטיסטיקות המקוריות לקבוצה L'2: P3 ו- 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

טבלה 39

סיכום התוצאות:

- לפי התזה של MBBK, קבוצה EM3 שלא עברה אופטימיזציה לגבי הכינויים, לא היתה צריכה להראות כאופטימום תחת הווריאציות. אבל בדיוק ההפך מזה קרה: לפי P4 קבוצה EM3 נראית כאופטימום לעומת התוצאה המקבילה לגבי L'2.

- התוצאות עבור A דומות, כמצופה, לתוצאות עבור EM3. גם כאן מוזר הדבר ש- A נראית כאופטימום לעומת התוצאה המקבילה לגבי L'2.

- לעומת זאת, לפי אותה תזה של MBBK: קבוצה B, שבהכרח ביחס לכינויים שבה נעשתה האופטימיזציה, צריכה להיראות כאופטימום חד תחת הווריאציות. והנה, מתברר שהתחלפו היוצרות, ושום אופטימום לא נתקבל עבור קבוצה B!

(ד) הבה נשוב ונבדוק, והפעם לפי סטטיסטיקות-r:

	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

טבלה 40

סיכום התוצאות:

- בעקבות מבחן הפרמוטציות התוצאות עבור EM3 ו- A לא השתנו באופן מהותי.
- לעומת זאת עבור B התוצאה r4 שונה מאד מהתוצאה עבור P4: מס' התוצאות שהתקלקלו עלה ב- 43%.
- עם כל זה, התוצאות ממשיכות להיות מפתיעות:

- דווקא עבור קבוצה B שבה היתה אמורה להתבצע האופטימיזציה, מספר השיפורים ב-r4 גדול מאשר עבור EM3, הקבוצה שאין חשש אופטימיזציה לגבי כינוייה, וכפול מאשר לגבי A.
- ולכן גם קורה, שעבור B, מספר השיפורים ב-r4 הוא 22, לעומת התוצאה של 4 שיפורים המוצגת בטבלאות 5-10 של MBBK. זו תוצאה מפתיעה לפי המודל של MBBK: L'2 היא האיחוד של A ו- B, ולכן היא מכילה קבוצת כינויים שלא עברה אופטימיזציה. אנו מצפים לכך, שחלק "אינרטי" זה יתרום לאיזון היחס בין שיפורים לקילקולים. אבל ההפך הוא שקרה: מספר השיפורים עבור B, שבה אמורה להתרכז כל האופטימיזציה, גדול **פי 5.5** מאשר עבור L'2.

(ה) כל התוצאות שקבלנו בסעיפים הקודמים מהוות הפרכה גמורה לתזה של MBBK: אין שום קשר בין תוצאות הווריאציות לבין עניין ה"אופטימיזציה". אומנם, אין בזה כל חדש: הוכחנו כבר בפרק הקודם כי תוצאות "מחקר הווריאציות" הן תולדה של "תפירת" הווריאציות. גם כאן אנו רואים את תוצאות ה"תפירה": MBBK "תפרו" את הווריאציות כדי להשיג מינימום בשיפורים עבור r4 ב- L'2 [ובזה השיגו מינימום בשיפורים עבור

min(r1-r4)], אבל לא "תפרו" אותן ביחס ל-EM3, ל-A או ל- B. אוסף הווריאציות ה"תפור", שהוא אוסף לא תקין, מוביל לתוצאות משונות כאשר מיישמים את "מחקר הווריאציות" על EM3, על A או על B.

אבל, יש מקום לעיין בתוצאות מבחן הפרמוטציות עבור EM3, A ו- B עיון נוסף. במדגמים המבוססים על EM3, על A ועל B ישנם מקרים רבים בהם אין לאישיות כינוי או תאריך. במקרה כזה אין שום תרומה של זוגות "כינוי-תאריך" למידגם עצמו, אך ישנה השפעה עקיפה דרך מבחן הפרמוטציות. MBBK טוענים שזה גורם "רעש רנדומלי" [16] ולכן הם מורידים את הנתונים שלא משתתפים בזוגות "כינוי-תאריך" במידגם עצמו.

עבור A ו- B עניין זה מגיע לקיצוניות, כאשר רק לפחות ממחצית האישים יש לפחות זוג "כינוי-תאריך" אחד. רצינו לבדוק מהי ההשפעה במקרה כה קיצוני, והורדנו את הנתונים שלא משתתפים בזוגות "כינוי-תאריך" במידגם עצמו.

תוצאות מבחן הפרמוטציות נראות עכשיו כך:

	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

טבלה 41

מן הטבלה ניתן ללמוד כי:

- חל שינוי ברור ב- r4 עבור B: ישנה עליה משמעותית (של 68%) במספר השיפורים, והתוצאות כמעט וחזרו להיות מאוזנות כמו עבור P4.
- עבור A: ישנה עליה מתונה יותר (של 27%) במספר השיפורים ב- r4, והתוצאות קרובות לערכן ב- P4.

בזה נעלמו השרידים האחרונים לאופטימום ב-B, הקבוצה בה היתה אמורה להתרחש ה"אופטימיזציה".

התלות החזקה של תוצאות הווריאציות במקרה זה ב"רעש", יחד עם תוצאות הניסוי בסעיף **(ב)** דלעיל, מחזקים את הרושם, שתוצאות הווריאציות של MBBK עבור r4 הן אנומליה הנובעת מתכונה זו או אחרת של מבחן הפרמוטציות , ללא קשר לקיומה או לאי קיומה של אופטימיזציה על הכינויים.

<u>לסיכום:</u>

(1). סיכום הניסויים בפרק ד (טבלה 32), יחד עם הניסויים בפרק זה, מראים בעליל כי אינם מתיישבים עם האפשרות, שהנחותיהם של MBBK נכונות, ושתוצאות מחקרם נובעות מאופטימיזציה של נתוני WRR.

(2). לעומת זאת ניתן להסביר את התוצאות על סמך ההנחה שהיה "tuning" של הווריאציות. הסבר זה נתמך גם על ידי ראיות ל"tuning" שהבאנו בפרקים הקודמים.

(3). הצבענו על תהליך האבולוציה של הווריאציות ושל צורת הצגת תוצאותיהן במגמה ברורה של שיפור התוצאה הרצויה ל- MBBK.

(4). נסינו להתחקות אחרי מקורה של "האופטימיזציה" שמדדו MBBK. לשם כך השתמשנו ברשימת השמות והכינויים של ד"ר עמנואל כבסיס. ואז התברר כי:

(i) דווקא השמות והכינויים שבחר עמנואל הם שהפגינו "אופטימיזציה".

(ii) לעומת זאת, השמות והכינויים שבחר הבלין ולא בחר עמנואל, לא הפגינו "אופטימיזציה".

(5). בבדיקה לפי סטטיסטיקות-r אותן העדיפו MBBK, נתברר כי גם בקבוצת השמות והכינויים שבחר הבלין ולא בחר עמנואל, נתגלתה לכאורה "אופטימיזציה" (אומנם, במידה **פחותה** מאשר בקבוצת השמות והכינויים שבחר עמנואל). בסופו של דבר התברר כי "תגלית" זו נבעה מ- artifact של מבחן הפרמוטציות במקרה זה. בסילוק הרעש – נעלמה "האופטימיזציה".

(6). כל זה מביא אותנו לכלל מחשבה, כי התוצאה של MBBK עבור r4 היא בסך הכל אנומליה שנתקבלה מתכונה זו או אחרת של מבחן הפרמוטציות, ללא קשר לקיומה או לאי קיומה של אופטימיזציה על הכינויים. ואם נכון הדבר, נמצא שה-"tuning" של הוריאציות נעשה בצורה כה שלומיאלית, כך שתוצאות הוריאציות תלויות תלות חזקה בתכונות מסוימות של מבחן הפרמוטציות, ותו לא.

נספח

<u>לפרק א:</u>

1. **לסעיף 1**: לדוגמא: בפרה-פרינט הראשון (86), אנו מדגישים את חשיבותם של <u>שני</u> מרכיבים במפגש הגיאומטרי בין שני ELSs: שכל אחד מן ה-ELSs יהיה ממוקד על פני הטבלה (או הגליל) הדו-ממדי, כלומר יהיה בעל "small localization parameter" (f קטן), ושהם יהיו קרובים זה לזה (l קטן). ראה עמ' 8-9 ו 29-30 שם. ואילו MBBK מתעלמים מכך בוריאציות המוצגות בטבלה 5.

2. **לסעיף 9**: הבה ונמנה את הסטטיסטיקות השונות בהן אפשר למדוד את תוצאות הוריאציות.

(א) ראשית, נמנה את האפשרויות בסטטיסטיקות-P. ברירת המחדל היא min(P1-P2), וכבר הצבענו על כך בגוף המאמר (תחילת פרק **ג**) כי מוזר הדבר ש-MBBK התעלמו מבחירה טבעית זו.

חוץ מזה, יש לנו 4 הסטטיסטיקות הידועות: P1, P2, P3, P4.

MBBK השתמשו גם ב- min(r1-r4), באותה המידה יכלו MBBK להשתמש גם סטטיסטיקה המקבילה ב- min(P1-P4).

סך הכל יש לנו כאן 6 סטטיסטיקות.

(ב) בסטטיסטיקות-r יש לנו כנגדן 6 סטטיסטיקות: r1, r2, r3, r4, min(r1-r2), min(r1-r4).

(ג) מלבד זאת, פרופ' בר-הלל השתמשה לצורך מחקר הוריאציות בעוד שתי סטטיסטיקות ש- WRR לא השתמשו בהן מעולם, כפי שמציין פרופ' אומן במכתבו אליה [9].

(ד) ומלבד זאת, השתמש מקי בדו"ח הראשון שלו [13], לצורך בדיקת "robustness" (בדיקה שהיא אחת ממטרותיו המוצהרות של "מחקר הוריאציות") בעוד שלוש סטטיסטיקות ש- WRR לא השתמשו בהן מעולם.

לכן, גם אם נסתפק במה שידוע לנו נגיע ל- 17 סטטיסטיקות אפשריות לכל מידגם. וכיוון ש-MBBK לא הקפידו לבחור אותן הסטטיסטיקות עבור שני המידגמים, הרי לפנינו:

$N=2^{34}$ אפשרויות, שהן יותר מ- 17,000,000,000 אפשרויות. מתוך מספר עצום זה, בחרו MBBK ארבע סטטיסטיקות מסוימות, שתים עבור המידגם הראשון, ושתים עבור המידגם השני.

איננו טוענים שכל הצירופים סבירים באותה המידה. קשה לדעת כמה סיפורים סבירים יכלו MBBK להמציא כדי להצדיק צירופים אפשריים מתוך 17,000,000,000 אפשרויות. אבל, דווקא MBBK מספרים לנו [22], שיש להם יכולת כבירה לייצר סיפורים מסוג זה, ועבורם מרחב הסיפורים הוא עצום וכמעט בלתי מוגבל.

הכרת תודה

אנו מודים בזה מקרב לב לד"ר שלום סרברניק על שתרם תרומה חשובה ליצירת מאמר זה. תודתנו נתונה ליואב רוזנברג וליעקב רוזנברג אשר בתוכנה שלהם השתמשנו לביצוע הניסויים המתוארים כאן.

ביבליוגרפיה

1. מאמרנו:

Witztum, D., Rips, E. and Rosenberg, Y. (1994). *Equidistant Letter Sequences in the Book of Genesis.* Statist. Sci. 9 No. 3 429-438.

2. מאמרם של MBBK:

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. ויצטום, ד. (התשס"א). על מדע ועל פרודיה: הפרכה גמורה של הטענה המרכזית של MBBK.

4. המאמר

Bar-Hillel, M., Bar-Natan, D. and McKay, B. D. (1998). *Torah codes: puzzle and solution.* Chance 11 No. 2 13-19.

5. הפרסום

McKay, B. D. (April 1997). *Equidistant letter sequences in Genesis – A Report* (draft).

6. התגובה

Bar-Hillel, M., Bar-Natan, D. and McKay, B. D. (1998). Reply, Chance 11 No. 4 50-51.

7. ויצטום, ד. (התש"ס). על המבחן הסטטיסטי שהתפרסם במאמרנו ב- Statistical Science (חלק ב).

8. ויצטום, ד. (התשס"א). הוכחה סטטיסטית חדשה לקיום צופן בספר בראשית.

9. המכתב

Aumann, R. J. (1997). A letter to Maya Bar-Hillel, dated 17 Jan.

10. ויצטום, ד. (התשס"א). על ה"וריאציות" של MBBK בעניין התאריכים. בהכנה. [פורסם לבסוף בשם: על בחירת התאריכים למדגמים של WRR].

11. הסקירה

Gans, H. J. (2000). *A Primer on the Torah Codes Controversy for Laymen.*

12. Private communication.

13. הפרסום

McKay, B. D. (Feb. 1997). *Equidistant letter sequences in Genesis – A Report* (draft).

14. התגובה

Witztum, D., Rips, E. (1998). Reply: Choice of Choices. Chance 11 No. 4 48-49.

15. התדפיס

Witztum, D., Rips, E. and Rosenberg, Y. (1986). *Equidistant letter sequences in the Book of Genesis.* Preprint.

16. המאמר

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. בר-הלל, מ. ובר-נתן ד. (1996). מכתב לפרופ' אומן, מתאריך 27 בנובמבר. שאלה מס' 5. נמצא בתוך: מסמך 2 (1997). בר-הלל ובר-נתן שואלים – ויצטום וריפס משיבים.

18. המכתב

Kalai, G. (1997). A letter to R. Aumann, dated 11 Nov 97.

19. המכתבים

Kalai, G. (1997). Letters to R. Aumann, dated 11 and 12 Nov 97.

20. ויצטום, ד. (התשנ"ח). על הרמז בדילוג השווה: מדידה חדשה של מדגם גדולי התורה. בדד, כתב-עת לעניני תורה ומדע, חוב' 7, הוצאת אוניברסיטת בר-אילן.

21. הרצאה

Bar-Hillel, M (1997). A lecture at the meeting of the Center for Rationality and Interactive Decision Theory at Zarka Ma'in.

22. המאמר

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.