

## פיתוח מדדים לבקרת איכות לציוני מבחנים המועברים תכופות

אבי אללוף  
טוני גוטנרג  
מיכל באומר  
סיגל לוי  
מרינה פרונטון  
קרן רודד

מאי 2014



מרכז ארצי לבחינות ולהערכה (ע"ר)

NATIONAL INSTITUTE FOR TESTING & EVALUATION

المركز القطري للامتحانات والتقييم

מיסודן של האוניברסיטאות בישראל

**דוח מרכז 400**  
**ISBN:978-965-502-181-3**

## **פיתוח מדדי בקרת איכות לציוני מבחנים המועברים תכופות**

אבי אללוף<sup>1</sup>, טוני גוטנטג<sup>1</sup>, מיכל באומר<sup>1</sup>, סיגל לוי<sup>2</sup>, מרינה פרונטון<sup>1</sup> וקרן רווד<sup>1</sup>

**מאי 2014**

המחקר קיבל מימון מקרן המרכז הארצי לבחינות ולהערכה 2012-2013

---

1 מאל"ו

2 המכללה האקדמית תל-אביב - יפו

צוות המחקר מודה

לפרופ' משה פולק, האוניברסיטה העברית, החוג לסטטיסטיקה, על ייעוצו ותרומתו למחקר  
לנדב פודולר שהצטרף לצוות לאחרונה (במקום קרן רווד) על הערותיו לדוח

## תוכן עניינים

|         |                                                                              |
|---------|------------------------------------------------------------------------------|
| 4.....  | מבוא                                                                         |
| 5.....  | א. בקרת איכות לציוני מבחנים המועברים לקבוצות אוכלוסיה גדולות במועדים מוגדרים |
| 7.....  | ב. בקרת איכות על ציוני מבחנים המועברים לקבוצות אוכלוסיה קטנות במועדים רבים   |
| 10..... | שיטה                                                                         |
| 10..... | משתתפים ונתונים                                                              |
| 10..... | כלים                                                                         |
| 10..... | הליך                                                                         |
| 12..... | תוצאות                                                                       |
| 27..... | דיון                                                                         |
| 32..... | מקורות                                                                       |
| 33..... | נספח 1 – שיטות לבקרת איכות                                                   |
| 34..... | נספח 2 – התפלגות נוסחי הבחינה בכול חודש באחוזים                              |
| 35..... | נספח 3 – אופן חישוב הגבולות                                                  |
| 36..... | נספח 4 – "מחוץ לשליטה"                                                       |
| 37..... | נספח 5 – טבלאות בקרה לכל נוסח בנפרד                                          |
| 46..... | נספח 6 – מקדמי מתאם רגרסיה ליניארית                                          |

## תקציר

הליכי בקרת איכות (QC – Quality Control) עבור ציוני מבחנים ניתן לחלק לשני סוגים: QC עבור בחינות המועברות לקבוצות אוכלוסייה גדולות במספר מצומצם של תאריכי בחינה תוך שימוש במספר נוסחים קטן, ו-QC עבור בחינות המועברות לקבוצות אוכלוסייה קטנות במספר רב של תאריכי בחינה תוך שימוש במספר נוסחים גדול. מחקר זה עוסק בסוג השני, להלן: בקרת איכות על מבחנים המועברים תכופות. בהינתן שעולם הבחינות מתקדם בקצב מהיר לכיוון זה, למחקר בתחום חשיבות רבה.

### מטרה ונתונים

מטרת המחקר הייתה לפתח מדדי QC לבחינה זו בפרט ועבור בחינות המועברות תכופות באופן כללי. המחקר השתמש בנתונים מ-24,548 נבחנים שניגשו לבחינה ממוחשבת (מימ"ד) במהלך ארבע שנים (2008 – 2012).

### מתודולוגיה, מערך ותוצאות

ראשית, נבדקה העברת הבחינה: בדיקה של **משקולות דגימת הנוסחים** מצאה כי הדגימה עבדה במתוכנן; ניתוח באמצעות מבחן חי-בריבוע מצא **שסדר הפרקים בכל נוסח** היה מקרי כמתוכנן. בשלב הבא נחקרו פריטי הבחינה. מבחן  $t$  מזווג מצא **יציבות קושי פריטים** מעבר לזמן. מספר פריטים שנמצאו יציבים פחות נבדקו לעומק, אך לא נמצאה סיבה עקבית לכך. **שקילות הנוסחים** נבחנו באמצעות ניתוח שונות. נמצא הבדל מובהק בין הנוסחים, אך גודל האפקט היה קטן. נוסח בחינה מסוים אופיין כבעייתי, ממצא שתמך בהחלטה שקדמה למחקר להפסיק את השימוש בו. לבסוף, נבחנו הציונים עצמם. **יציבות הציונים לאורך זמן** נוטרה באמצעות טבלאות בקרה מסוג שוהארט. חלק ניכר מהשונות הכללית הוסברה באמצעות מגדר, הארכת זמן ומוסד אקדמי בו נבחנו המשתתפים. כאשר שורטטו טבלאות בקרה מסוג שוהארט עבור שאריות הניבוי נמצאו חריגות מועטות מאד לאורך השנים. ציוני **הנבחנים חוזרים** נמצאו דומים בין העברות.

### מסקנות וכוונות לעתיד

יישום של תהליכי בקרת איכות יכול לסייע לזיהוי מהיר ומדויק של בעיות בנוסחי בחינה קודמים ועכשוויים; המחקר בנושא ממשיך כדי לשפר את הבדיקות שנעשו ולפתח שיטות חדשות. מטרה עתידית הינה ליצור שגרת בדיקות אוטומטית ככל האפשר כדי לנטר בעיות בזמן אמת. בנוסף, יש כוונה לערוך בקרת איכות בפורמט דומה, עם ההתאמות הנדרשות, גם על בחינות אחרות המועברות באופנות מתמשכות.

**מילות מפתח:** בקרת איכות, ציינון, מבחנים מבוססי מחשב; הערכה בקנה מידה גדול; מבחנים המועברים תכופות.

## **Developing Quality Control Measures for Scores of Tests with Continuous Administration Mode**

Quality Control (QC) procedures for test scores can be divided into two types: QC on tests that are administered to large population groups on few administration dates using a small set of test forms, and QC on tests administered to small population groups on many administration dates using a wide array of test forms. This research relates to the latter, that is, tests that are continuously administered. Given that the world of testing is rapidly headed in this direction, research in this field is highly relevant.

**Objectives:** The study used data pertaining to 24,548 examinees who took an online test (MEIMAD) over a period of four years (2008-2012). The goal is to develop QC measures for this test in particular and for continuously administered tests in general.

**Methodology, Design & Results:** The study first looked at test administration: A descriptive examination of **a weighted sampling of test forms** found that the sampling worked as planned; an analysis using a chi-square test found that **the order of sections in each test form** was random, as planned. Next, we looked at the test items: An analysis using paired t-tests found **stability in item difficulty** over time. Several test items that were less stable were examined in depth, but no consistent reason for their lack of stability could be found. ANOVA was used to examine **test form equivalence**. A significant difference was found between test forms, but effect size was small. One particular test form was found to be problematic, supporting a decision (made before this study was conducted) to stop using it. **Stability of scores over time** was tracked using a Shewhart control chart; an undesirably high level of deviation was found. A substantial part of the overall variance was explained well by examinee demographics and other variables. When plotting a Shewhart control chart for the predictive residuals, very few deviations were found over the years in question. Finally, data was extracted with regard to **repeat examinees**, which will allow for further QC.

**Conclusions & Future Directions:** Application of QC methods can facilitate fast and accurate identification of problems in past and current test forms; research on this subject is conducted in order to improve QC methods and develop new ones. An important goal is to put these methods into place – to create, as much as possible, an automatic, routine QC procedure in order to neutralize problems in real time. Moreover, thought is being given to performing QC in a similar format, within the required parameters, on other continuously administered tests.

**Keywords:** quality control; scoring; computer-based testing; large-scale assessment; continuously administered tests.

## פיתוח מדדים לבקרת איכות לציוני מבחנים המועברים תכופות

### מבוא

בקרת איכות על ציוני מבחנים הינה הליך סטטיסטי פורמאלי המתוכנן להבטיח שסטנדרטי איכות מושגים במהלך ציינון, כיול ודיווח על תוצאות המבחן (Allalouf, 2007). תהליכי הציינון-כיול-דיווח הם מורכבים ורב שלביים – כאשר בכל שלב עלולה להתרחש טעות. מכיוון שבמבחנים עתירי סיכון (High Stakes) מחיר הטעויות הוא גבוה או אף גבוה מאוד (לנבחנים, למשתמשים [מוסדות הלימוד] ולעורכי המבחנים), יש צורך בבקרת איכות על תהליכי המדידה ועל הציונים המתקבלים בפרט. במחקר זה עסקנו בבקרת איכות על ציוני מבחנים המועברים לקבוצות אוכלוסייה קטנות במועדים רבים ובנוסחים מגוונים (אופנות העברה רצופה או מתמשכת), תחום מאתגר וחדשני אליו משתנה עולם הבחינות כיום.

בקרת איכות במבחנים הינה תחום חדש, הנשען על שיטות שפותחו לצורך בקרת איכות על מוצרים תעשייתיים לפני כמאה שנה. מבשרי פיתוח שיטות סטטיסטיות לבקרת איכות בתעשייה היה שוהארט (Shewhart, 1931, 1939) וממשיכו דמינג (Deming, 1982). מקור חשוב ועדכני לשיטות סטטיסטיות לבקרת איכות הוא מונטגומרי (Montgomery, 2009). בקרת איכות מפותחת וחשובה במיוחד בדיסציפלינות שבהן לטעויות מחיר כבד, ולכן בקרת איכות בהם הינה קריטית. זו הסיבה ששיטות בקרת איכות נפוצות ומפותחות מאוד בתחום הרפואה והתעופה, שם המחיר של טעות באבחנה רפואית או בניתוח רפואי, כמו גם המחיר של טעות בניווט או בנחיתה של מטוס במזג אוויר סוער – עלול להיות גבוה מאוד. כלים לבקרת איכות נפוצים גם בתחום של בנייה אזרחית (בתים, גשרים וכו') ובתחום פיתוח חומרה ותוכנה. בקרת איכות בתעשייה מתבצעת לרוב על דגימות שיטתיות הנלקחות תוך כדי תהליך הייצור.

כמובן, שבקרת האיכות הנערכת על מבחנים שונה מאוד באופייה מבקרת איכות על מוצר תעשייתי. אחד ההבדלים קשור לכך שמוצר תעשייתי הוא מוחשי ולפחות חלק מהפגמים, אם יש כאלה, נראים לעין המתבונן ויתגלו לכן בטווח הקצר, לעיתים תוך כדי השימוש במכשיר. כשמדובר במבחנים חינוכיים ופסיכולוגיים, התוצרים הינם מסוג אחר, תוצרי תהליכי מדידה האמורים לייצג תכונות, יכולות וכישורים של האינדיבידואל שאינם מוחשיים. אלה, מבוטאים לרוב כתוצר מספרי: ציונים ומדדים סטטיסטיים. אם נפלה בהם שגיאה, מירב הסיכויים שהיא תתגלה מאוחר או לא תתגלה לעולם. בנוסף, תהליכי בקרת איכות על מוצרים נערכים לרוב על מדגם מקרב המוצרים, באופן רציף (בתכיפות גבוהה); בעוד שתהליכי בקרת איכות על מבחנים נערכים לרוב על אוכלוסיית הנבחנים כולה, בכל תקופת זמן קצובה.

בקרת איכות בתחום המדידה וההערכה הפסיכולוגית והחינוכית חיונית בשל מורכבות התהליך. ראשית, חישוב ציונים לבחינות הוא תהליך רב-שלבי ארוך: החל מקליטת נתונים, עובר דרך חישוב ציוני גלם, ניתוחי פריטים, כיול ומסתיים בחישוב ציוני תקן. כל אחד מהשלבים האלה נסמך על השלבים הקודמים וכולל בתוכו תתי-שלבים. שנית, הנחות סטטיסטיות המשמשות בסיס לחישובים אינן מתקיימות בהכרח באופן מושלם, דבר המוסיף מורכבות לתהליך. שלישית, התהליך אינו קבוע מראש, אלא כולל צמתי החלטה המחייבים שיקול דעת.

בסיום תהליך ההערכה והמדידה יש בידינו ציונים לכל נבחן. **התפקיד העיקרי של בקרת האיכות הוא לאשר שהציונים הם סבירים וניתן לדווח אותם.** בנוסף, בקרת האיכות יכולה להנחות אותנו כיצד לשפר את התהליך בבחינות עתידיות.

מוסדות האמונים על עריכת בחינות מודעים לחשיבות תהליך בקרה מסודר על הציונים. שיטות בקרה בסיסיות, כמו חישוב כפול של ציונים (חישוב בשתי דרכים והשוואה בין התוצאות), או שימוש ברשימת בדיקות (Check List) כדי לוודא שלא נשכח אף שלב בתהליך, הן שיטות בסיסיות ופשוטות יחסית לעשות כן. בנוסף לשיטות אלה, קיימות שיטות מקובלות מורכבות יותר לבקרת איכות של מבחנים. אנשי מקצוע מתחום הפסיכומטריקה מגלים בכך לאחרונה עניין גובר והולך, ומספר המחקרים והספרים העוסקים בכך עולה בהתמדה, למשל הספר שנערך על ידי ווילד ורמסואמי (Wild & Ramaswamy, 2008) שמאמרו מתארים כיצד ניתן להשתמש בשיטות לבקרת איכות שמשמשות בייצור ובשירותים גם בתעשיית המבחנים.

ניתן לחלק את בקרת האיכות על ציונים לשני סוגים עיקריים: (1) בקרת איכות עבור בחינות המועברות לקבוצות אוכלוסייה גדולות במספר מצומצם של תאריכי בחינה תוך שימוש במספר קטן של נוסחים הנקבעים מראש, ו-(2) בקרת איכות על בחינות המועברות לקבוצות אוכלוסייה קטנות במספר רב של תאריכי בחינה תוך שימוש במספר נוסחים גדול כאשר הנוסח הניתן לכל נבחן אינו בהכרח קבוע מראש. מחקר זה עוסק בסוג השני, שהוא מאתגר וחדשני, שכן עולם הבחינות משתנה לכיוון זה. כדי להבין מה ייחודי בסוג זה, נסקור תחילה את הסוג הראשון.

#### א. בקרת איכות לציוני מבחנים המועברים לקבוצות אוכלוסייה גדולות במועדים מוגדרים

בלוח 1 מוצגות שש בדיקות הכרחיות לבקרת איכות על ציונים לפי קולן וברנן (Kolen & Brennan, 2004).

##### לוח 1: בקרת איכות על ציונים (Kolen & Brennan 2004)

1. Check that the administration conditions are followed properly
2. Check that the answer key is correctly specified
3. Check that the items appear as intended
4. Check that the equating procedures specified are followed correctly
5. Check that the score distribution and score statistics are consistent with those observed in the past
6. Check that the correct conversion table or equation is used with the operational scoring

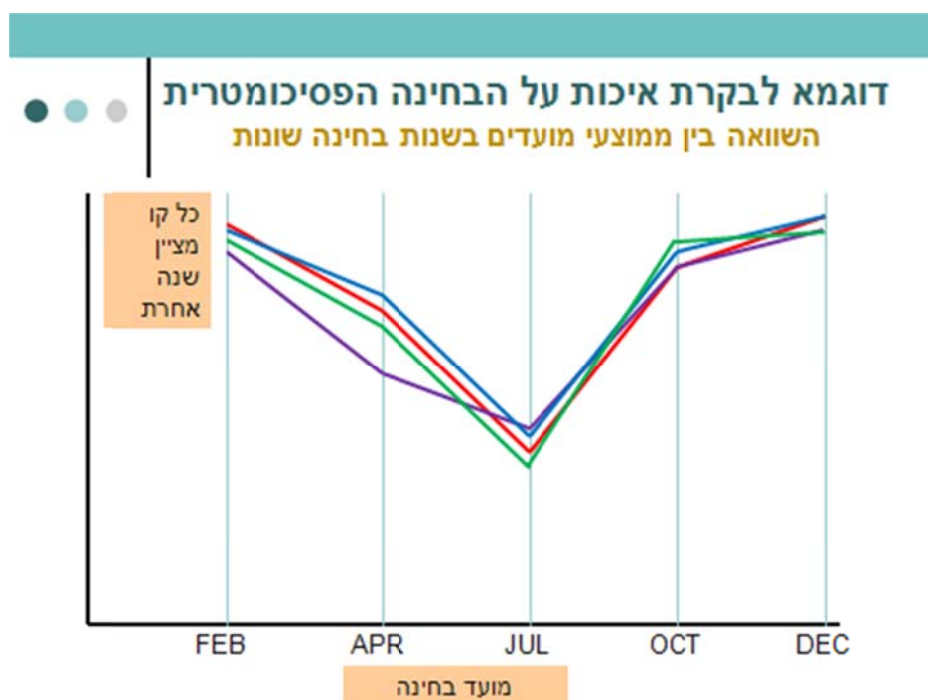
את הבדיקות המופיעות בלוח 1 ניתן לחלק לשלוש: (1) בדיקה שהמבחנים הועברו באופן סטנדרטי [בדיקה מס' 1], (2) בדיקות שלא חלה טעות בחישוב ובנתונים [בדיקות 2, 3, 4, 6] ו-(3) בדיקה שהתפלגות הציונים שחושבו מתיישבת עם התפלגות מקבילה בעבר [בדיקה 5]. הבדיקה האחרונה היא הרלוונטית ביותר לענייננו. בעוד ששאר הבדיקות (בהנחה שהעברה הייתה סטנדרטית) רק מוודאות שלא נפלה טעות, בדיקה 5 משווה את שהתקבל לצפוי. אם כל הבדיקות יוצאות "בסדר", אזי זו הבדיקה היחידה שיכולה להתריע על בעיה – בגלל שהציונים לא מתיישבים עם הצפוי.

השאלה היא, כמובן, מה צפוי? אללוף (Allalouf, 2007) סוקר כמה תחזיות אפשריות מהתפלגות הציונים של קבוצת נבחנים מסוימת (האפשרות הראשונה מתוכן זהה לבדיקה החמישית של קולן וברנן [Kolen & Brennan, 2004]): (1)

השוואה לזמן ההעברה המקביל אשתקד, (2) ניתוח השפעת נתונים דמוגרפיים (גיל, מגדר ועוד) של הנבחרים על ההישגים בבחינה על סמך ידע קודם אודות אופי ההשפעה, (3) בחינת שיפור ההתפלגות בקרב נבחרים שזו בחינתם השנייה ו-(4) השוואת ציונים בין חלקי הבחינה. ארבע האפשרויות האלה מתאימות היטב למצב של מבחנים המועברים לקבוצות גדולות של נבחרים במועדים קבועים ומוגדרים מדי שנה.

במרכז הארצי מתנהלת בקרה שגרתית על בחינות הנייר והעיפרון (Allalouf, 1999). הבקרה המפותחת ביותר מתקיימת על הבחינה הפסיכומטרית, בה קל יחסית לבצע בקרות – שכן כל נוסח בחינה מועבר לרוב לנבחרים רבים (כמה אלפים), ויש על הנבחרים נתונים דמוגרפיים רבים. ניתן לבחון הבדלים בהישגי הנבחרים בין מועדי בחינה שונים. כמו כן, מתבצעות בקרות באמצעות מספר משתנים עיקריים הנלקחים בחשבון – שפת הבחינה, מועד הבחינה, שיפור נבחרים חוזרים ("כפולים") והשוואה לציונים בתחומים אחרים; וכן מתבצעות בקרות באמצעות מספר משתני אוכלוסייה משניים – גיל, אחוז הנבחרים החוזרים, קמפוס הבחינה ומוסדות יעד אליהם ביקש הנבחן להעביר את ציוניו. יש לציין כי הבקרות שהוזכרו לעיל הן על התוצר הסופי – הציונים. בקרות אחרות מתבצעות על כל תהליך קליטת הנתונים וניתוח המבחן, למשל ניתוח פריטים וכיוול. לוח 2 ממחיש את אחת מהבקרות על ציוני הנבחרים בבחינה הפסיכומטרית – השוואת ממוצע הציונים במועדים השונים למועדים המקבילים בשנים הקודמות. ניתן לראות כי במועד אפריל (APR) הציון הממוצע באחת השנים נמוך מהציונים הממוצעים בשנים האחרות.

**לוח 2: המחשת בקרה על ציוני הבחינה הפסיכומטרית במרכז הארצי לבחינות השוואת ממוצע ציוני שנת הבחינה בהשוואה לשנות בחינה מקבילות קודמות**



## ב. בקרת איכות על ציוני מבחנים המועברים לקבוצות אוכלוסיה קטנות במועדים רבים

במבחנים המועברים תכופות (CAM – Continuous Administration Mode), שהם לרוב מבחנים ממוחשבים, קשה ולעתים בלתי אפשרי, לבצע בקרות בשיטות המקובלות. הסיבות העיקריות לכך: (1) מספר נבחנים קטן בכל העברה, (2) ריבוי נוסחים המועברים בכל שנת בחינות, (3) פרקי זמן קצרים בין זמני העברות המבחן ו-(4) רצון המערכת בדיווח ציונים מהיר הן למשתמשים (מוסדות לימוד) והן לנבחנים, מותיר זמן מועט לביצוע בקרות בזמן אמת לפני דיווח הציונים. עם זאת, גופים מקצועיים העורכים מבחנים גורליים (High Stake) – חייבים לבצע בקרות.

ון דויר (von Davier) מארגון ה-ETS (Educational Testing Service) סוקרת כמה מהשיטות הסטטיסטיות המקובלות בבקרת איכות אשר מתאימות לבקרת CAM (Lee & von Davier, 2013). בין השיטות: תרשימי בקרת איכות לחקר הנתונים לאורך זמן, סדרות עתיות (רגרסיות הרמוניות לנטרול עונתיות ו-Durbin-Watson test), תרשימי בקרה של סכום מצטבר (CUSUM), מודלי נקודת-שינוי (Change Point Model), מודלי מרקוב לאיתור שינוי פתאומי (HMMs), מודלים מעורבים משוקללים לאתר מבנים בנתונים ולניטור קשרים מבניים בין משתנים לאורך זמן, וניתוח שונות לניתוח מבנים בנתונים ולנטר קשרים מבניים בין משתנים לאורך זמן. נספח 1 מפרט כמה מהשיטות הנ"ל.

### תרשימי בקרת איכות

מטרת תרשימי בקרת איכות (QCC – Quality Control Charts), ובאופן ספציפי טבלאות בקרה של שוהארט (Shewhart Control Chart) השכיח והקל לפרשנות בינם, היא לספק מידע שיכול לחשוף פערים או תבניות סיסטמיות על ידי השוואת השונות המצופה לשונות הנצפית (Shewhart, 1931, 1939). תרשימים אלה מאפשרים לבדוק את השתנות המשתנה המבוקר לאורך ציר הזמן; למעשה, זוהי ההיסטוריה המצטברת של סטטיסטי המוצגת בתרשימים. טבלאות הבקרה השכיחות ביותר הינן M (ממוצע), S (סטיית תקן) ו-R (טווח).

תרשימי בקרת איכות משורטטים על פני שני ממדים: זמן (או מאורע) על הציר האופקי (ציר ה-X) ותצפית על הציר האנכי (ציר ה-Y). הגרף כולל שלושה קווים: קו אמצעי – קו הבסיס (Baseline) שהוא בדרך כלל מדד נטייה מרכזית או ערך מנובא; וגבול עליון ותחתון, המשורטטים כקו מרוסק ומשאירים מרחק שווה מעל ומתחת לקו האמצעי (בדרך כלל 3 סטיות תקן מעל ומתחת לקו הבסיס, בהתאמה), כפי שמוגדר להלן:

$$\text{Upper Control Limit} = \mu_x + 3 \sigma_x$$

$$\text{Baseline} = \mu_x$$

$$\text{Lower Control Limit} = \mu_x - 3 \sigma_x$$

התחשבות בגודל המדגם (כלומר חלוקה בשורש N) מובילה לשימוש ב"טעות תקן" במקום "סטיית תקן" (ר' נספח 3 לגרסה ששומשה במחקר הנוכחי).

סטיות מגבולות הבקרה, ו/תבנית שיטתית המופיעה בתרשימים, מוגדרים "מחוץ לשליטה" (Out of Control), ואילו תנודות בתוך גבולות הבקרה, כאשר דפוס הסטיות נראה רנדומאלי, מוגדר "בשליטה" (In Control). למעשה, נקודה הנופלת מחוץ

לגבולות היא כמו דחייה של השערת האפס בבדיקת השערות. מונטגומרי (Montgomery, 1985), בתוך: Schafer et al., (2011) הגדיר שישה מאפיינים מרכזיים, אשר אם אחד או יותר מהם מתרחש נאמר שהתהליך (או הערך) הוא "מחוץ לשליטה":

- (1) נקודה אחת או יותר היא מחוץ לגבול העליון או התחתון
- (2) לפחות שבע נקודות עוקבות הן באותו צד של קו הבסיס
- (3) שתיים מתוך שלוש נקודות עוקבות הן מחוץ לטווח ה- $2SD$
- (4) ארבע מתוך חמש נקודות עוקבות הן מחוץ לקו ה- $1SD$
- (5) כל דפוס מובחן שאינו רנדומאלי או שהינו שיטתי במובן כלשהו
- (6) נקודה אחת או יותר הקרובה לגבול העליון או התחתון

במאמרם מפרטים שפר ואח' (Schafer et al., 2011) שימושים שנעשו בטבלאות בקרה של שוהארט להערכה רחבת היקף של תכנית חינוכית ב-2011 על ידי המועצה הפסיכומטרית הלאומית של מרילנד. החוקרים מונים מדדים רבים ששימשו בבקרה זו, בינם:

- ממוצעים וסטיות התקן של הפריטים
- מתאמים בין פריטים
- ממוצעים וסטיות תקן של פרמטרי ההבחנה של הפריטים (IRT)
- הפרשים בין האשכול (צירוף כמה מבחנים) בעל הממוצע הגבוה ביותר והאשכול בעל הממוצע הנמוך ביותר, וכן הפרשים בין סטית התקן הגבוהה והנמוכה ביותר, בין טעות המדידה המינימאלית והמקסימאלית, בין הציונים הגולמיים המינימאליים והמקסימאליים ועוד
- בחינת גודל האפקט (Cohen's d) של ההפרשים דלעיל
- מהימנות פנימית (קרונבך אלפא)
- אחוז התלמידים שקיבלו ציוני קיצון מסך הנבחנים, ואחוז התלמידים שקיבלו ציוני קיצון מסך הנבחנים שנבחנו באשכול הספציפי הנבדק
- הפרשים בין האחוזונים הגבוהים ביותר לבין האחוזונים הנמוכים ביותר

**במרכז הארצי** בקורות מסוג זה עדיין אינן מפותחות דיין. קיימת שגרת בדיקה שנתית ותקופתית בה הציונים מנותחים ומושווים לשנים קודמות, אך טרם התבצע מחקר מקיף לפני המחקר המתואר להלן. עבודה ראשונית בתחום נעשתה לא מכבר על נתוני בחינת מימ"ד (אללוף ופרונטון, 2011) ששימשה כמושא המחקר גם במחקר הנוכחי (ר' פרק ה"שיטה"). נבדק ממוצע הציונים החודשי לכל אחד מששת נוסחי הבחינה בשנת 2011, ונבדקו כמה משתנים הקשורים לדגימה אקראית של הנוסחים השונים לנבחנים. המסקנה העיקרית הייתה כי דרוש מחקר רציני ומעמיק בנושא.

## מטרות המחקר

למחקר ארבע מטרות התלויות זו בזו:

- א. בקרת איכות על מבחנים מרובי העברות שכבר הועברו – איתור וחקירה לעומק של מצבים בעייתיים בעבר
  - ב. פיתוח שיטות לבקרת איכות מבחנים מרובי העברות
  - ג. ניסוח שגרת בדיקות עבור המבחנים הקיימים
  - ד. ניסוח שגרת בדיקות כוללת שתתאים גם למבחנים עתידיים (אמירנט – בחינה ממוחשבת באנגלית הצפויה להיות תפעולית בזמן הקרוב, ובעתיד הרחוק יותר הבחינה הפסיכומטרית הממוחשבת)
- כדי להשיג מטרות אלה נעשה שימוש בשיטות שתוארו לעיל בסקירת הספרות, ובשיטות נוספות שפותחו תוך כדי המחקר.

## שיטה

### משתתפים ונתונים

בסיס הנתונים של המרכז הארצי עמד לרשות החוקרים. בבסיס נתונים זה קיימים הציונים של כל הנבחנים בבחינות של המרכז הארצי החל משנת 1983. במימ"ד נבחנו בין השנים 2008 (אפריל) עד 2012 (מרץ) 24,548 נבחנים.

### כלים

להלן תיאור קצר של בחינת המימ"ד (המרכז הארצי לבחינות ולהערכה, 2013):

| מספר נוסחים | שנת בחינות | נבחנים                | תוכן ומבנה     | מטרות                  | בחינה |
|-------------|------------|-----------------------|----------------|------------------------|-------|
| 2           | 2008       | מועברת החל מ-1990.    | שלושה תחומים:  | מיון וקבלה של המועמדים | מימ"ד |
| 4           | 2009       | בשנים האחרונות ניגשים | עברית, מתמטיקה | למסלולי הלימוד השונים  |       |
| 5           | 2010       | לבחינה כ-5,000 נבחנים | ואנגלית.       | במכונות הקדם-אקדמיות.  |       |
| 6           | 2011       | בשנה.                 |                |                        |       |

מימ"ד היא בחינת ידע למועמדים למכונות הקדם-אקדמיות בישראל (באומר, גפני, טורוול ושבדי, 2010); המרכז הארצי לבחינות ולהערכה, 2013). מוסדות הלימוד משתמשים בציון בחינת המימ"ד ובציוני בחינות הבגרות ככלי למיון המועמדים למסלולי הלימוד השונים במכונות הקדם-אקדמיות.

הבחינה מודדת ידע בשלושה תחומים: (א) **עברית**. בתחום זה נבדקים העושר הלשוני והיכולת המילולית הבאים לידי ביטוי בלימודים אקדמיים; (ב) **מתמטיקה**. בתחום זה נבדקות היכולת להשתמש במספרים ובמונחים מתמטיים והיכולת לנתח נתונים המוצגים בצורות שונות כגון טבלאות ותרשימים; ו-(ג) **אנגלית**. בתחום זה נבדקת שליטה בשפה האנגלית, כפי שהיא באה לידי ביטוי בקריאת קטעים ומשפטים ובהבנתם.

**מבנה:** הבחינה מורכבת משמונה פרקים – שניים עד ארבעה מכל תחום, כך שישנם שישה פרקים תפעוליים (שנים מכל תחום) ושני פרקים ניסיוניים מאחד משלושת התחומים. בראש כל פרק מצוינים מספר השאלות בו ומשך הזמן המוקצב לפתרונן. משך הבחינה כולה כשלוש שעות וחצי. השאלות בבחינה הן שאלות ברירה: לכל שאלה מוצעות ארבע תשובות, מתוכן על הנבחנים לבחור את התשובה הנכונה ביותר.

**הפעלת הבחינה על ידי המרכז הארצי:** החל ממאי 2008 הועברה הבחינה לפורמט ממוחשב (באומר, גפני, טורוול ושבדי, 2010). העברת הבחינה מתבצעת במוסדות לימוד אקדמיים ברחבי הארץ על ידי נאמן בחינה מטעם המרכז הארצי.

### הליך

המשתתפים והנתונים שעברו בקרה והשיטה שבה נערכה הבקרה הינם:

- דגימת נוסחים.** בכל מוקד בחינה מועברים כמה נוסחים, כאשר התוכנה דוגמת לכל נבחן נוסח בחינה. משקלו של כל נוסח בדגימה אינו בהכרח שווה, ונמצא בשליטת מפעילי הבחינה. שליטה זו מאפשרת להבטיח שכל נוסח יועבר למספר הנבחנים הדרוש. התבצע באמצעות מבחן חי בריבוע (השוואת ההתפלגות הרצויה להתפלגות המצויה).
- סדר פרקים רנדומאלי.** לאחר שנדגם נוסח בחינה מסוים, התוכנה מציגה את שמונת הפרקים המרכיבים אותו בסדר אקראי. טכניקה זו מאפשרת להפחית למינימום את הסיכוי ששני נבחנים היושבים זה לצד זה יבחנו באותו פרק באותו הזמן. הבדיקה התבצעה באמצעות מבחן חי בריבוע (לבחינה שהתפלגות ההקצאה למיקומים שונים של הפרקים רנדומאלית).
- יציבות קושי הפריטים.** הפריטים הנכללים בבחינה אמורים להיות יציבים, דהיינו, לשמור על תכונותיהם הפסיכומטריות מעבר להעברות השונות. התבצע באמצעות בחינת השתנות קושי הפריטים מעבר לשנים באמצעות מבחן t מזווג.

4. **שקילות נוסחי הבחינה.** הנוסחים עוברים כיוול לפני העברתם. הם אמורים להיות שקולים (כלומר – הציון לא מושפע מהנוסח הספציפי בו האדם נבחן). ניתוח שונות שימש לבחינת שקילות הנוסחים. הניתוח נעשה בנפרד על כל משתנה תלוי: הציון בחלק העברית, הציון בחלק המתמטיקה, הציון בחלק האנגלית וציון הבחינה הכללי, כאשר המשתנה הבלתי-תלוי היה נוסח הבחינה.

5. **יציבות ציונים לאורך זמן.** בחינה האם הציונים יציבים מעבר לזמן.

**א. טבלאות בקרה** מסוג שוהארט לבחינת השתנות ציוני תחומי הבחינה וציון הבחינה הכללי מעבר לזמן.

**ב.** ניבוי ציון הבחינה הכללי באמצעות **רגרסיה ליניארית**.

**ג.** טבלאות בקרה מסוג **שוהארט עבור שאריות** ציון הבחינה הכללי.

6. **הישגי נבחנים חוזרים.** נבדק אחוז הנבחנים החוזרים לפי מספר העברה, הציון הממוצע של נבחנים חוזרים לפי מספר העברה, שיפור הציון הממוצע של נבחנים חוזרים מהעברה להעברה, וההפרש הממוצע בימים בין זמני ההעברות. כמו כן, נערך מעקב אחר נוסחי הבחינה שנבחנים חוזרים מקבלים, על מנת לבחון שנבחנים חוזרים אכן מקבלים נוסח בחינה שונה בכל העברה כמתוכנן.

לפני הפעלת כלי הבקרה הופקה **סטטיסטיקה תיאורית** של ציוני תחומי הבחינה (עברית, מתמטיקה ואנגלית) וציון הבחינה הכללי. כמו כן, נבדק **הקשר** בין ציוני תחומי הבחינה וציון הבחינה הכללי באמצעות מתאם פירסון.

## תוצאות

### 1. סטטיסטיקה תיאורית התפלגות

בטבלה 1 מתוארת התפלגות מספר הנבדקים לפי חודשים ושנים. ההתפלגות המלאה הכוללת אחוז נוסחי בחינה בכל חודש מצויה בנספח 2. שנת בחינות במימ"ד מתחילה בחודש אפריל ומסתיימת בחודש מרץ של השנה העוקבת, לדוגמא שנת הבחינות 2008 החלה באפריל 2008 והסתיימה במרץ 2009.

**טבלה 1: התפלגות מספר הנבחנים לפי חודשים ושנות בחינה**

| חודש/שנה | 2008 | 2009 | 2010 | 2011 |
|----------|------|------|------|------|
| אפריל    | 0    | 0    | 36   | 22   |
| מאי      | 78   | 186  | 128  | 77   |
| יוני     | 531  | 449  | 367  | 520  |
| יולי     | 866  | 987  | 618  | 564  |
| אוגוסט   | 1303 | 1358 | 1636 | 1244 |
| ספטמבר   | 1292 | 1353 | 846  | 1206 |
| אוקטובר  | 188  | 227  | 626  | 436  |
| נובמבר   | 0    | 216  | 123  | 95   |
| דצמבר    | 366  | 491  | 318  | 457  |
| ינואר    | 791  | 367  | 791  | 834  |
| פברואר   | 688  | 450  | 481  | 617  |
| מרץ      | 78   | 107  | 96   | 38   |

ניתן לראות שבעוד שיש חודשים בהם מספר הנבדקים בדרך כלל קטן מאוד, כגון אפריל ומאי, ישנם חודשים בהם מספר הנבחנים בדרך כלל גדול מאוד, כגון אוגוסט וספטמבר.

### סטטיסטיקה

בטבלה 2 מוצגים ממוצע וסטיית תקן של ציוני התחומים וציון הבחינה הכללי לפי שנות בחינה 2008-2011. ניתן לראות כי אין הבדלים משמעותיים בין השנים בסטטיסטיקה של הציונים המחושבים.

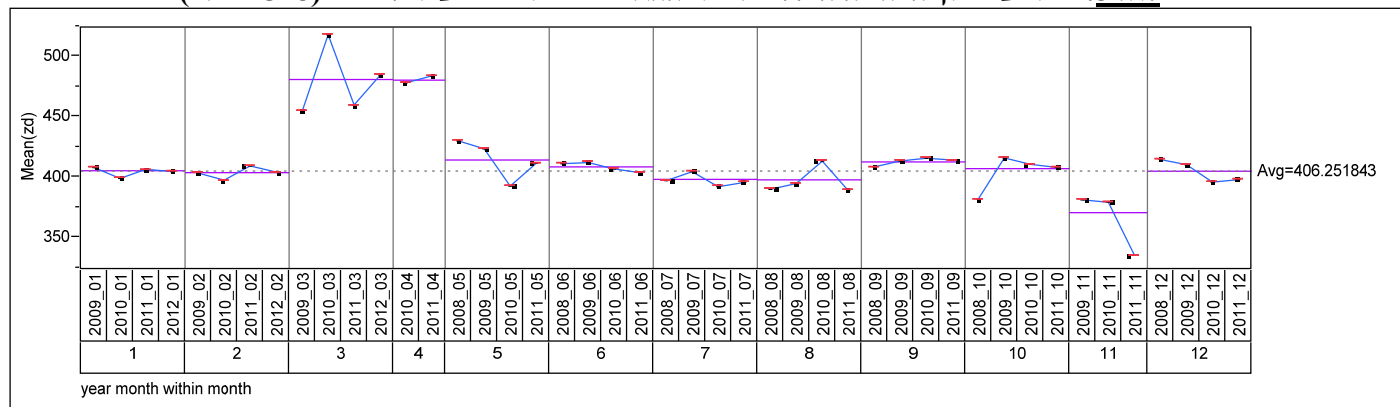
**טבלה 2: ממוצע (וסטיית תקן) של ציוני התחומים וציון הבחינה הכללי לפי שנות בחינה 2008-2011 וסך הכול מעבר לשנים**

| תחום/שנת בחינה    | 2008      | 2009      | 2010      | 2011      | סה"כ      |
|-------------------|-----------|-----------|-----------|-----------|-----------|
| N                 | 6,181     | 6,227     | 6,030     | 6,110     | 24,548    |
| עברית             | 80 (20)   | 81 (20)   | 81 (20)   | 80 (20)   | 80 (20)   |
| מתמטיקה           | 82 (21)   | 82 (22)   | 83 (21)   | 81 (21)   | 82 (21)   |
| אנגלית            | 81 (22)   | 83 (23)   | 82 (24)   | 82 (24)   | 82 (21)   |
| ציון הבחינה הכללי | 404 (105) | 408 (107) | 409 (107) | 403 (108) | 406 (107) |

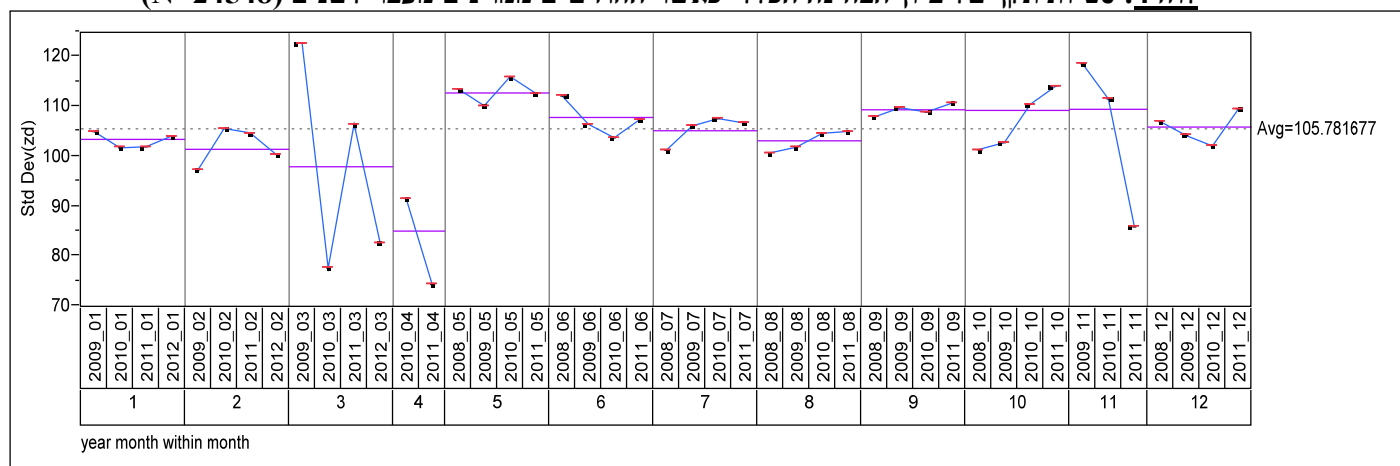
לוחות 3 ו-4 מציגים נתונים על אודות הממוצעים וסטיות התקן, בהתאמה, לפי חודשים. ניתן לראות שיש חודשים תנודתיים מאוד (למשל, מרץ) ואילו חודשים אחרים הנראים יציבים יותר (למשל, ינואר). לפי הממוצעים (לוח 3), נראה שלחודשים שונים מעבר לשנים יש ממוצעים דומים. לפי סטיות התקן (לוח 4), ניתן לראות שבחודשים עם מספר נבחנים גדול (ר' טבלה 1 – התפלגות מספר הנבחנים לפי חודשים ושנות בחינה) יש פחות תנודתיות בסטיות התקן בהשוואה לחודשים עם מספר

נבחנים קטן. בשילוב ממצאי הממוצעים וסטיות התקן, נראה שבחודשים בהם מספר הנבחנים גדול יש פחות תנודות בסטיות התקן וכן הממוצע דומה יותר לממוצע הכללי מעבר לתקופה, ולהפך לגבי חודשים בהם מספר הנבחנים קטן.

### לוח 3: ממוצעי ציון הבחינה הכללי כאשר החודשים ממוינים מעבר לשנים (N=24548)



### לוח 4: סטיות תקן של ציון הבחינה הכללי כאשר החודשים ממוינים מעבר לשנים (N=24548)



### מתאמים

בטבלה 3 מוצגים מתאמים בין חלקי הבחינה לציון הבחינה כללי לפי שנות בחינה 2008-2011, ומעבר לשנים.

### טבלה 3: מתאמים בין חלקי הבחינה וציון הבחינה הכללי לפי שנות בחינה 2008-2011 וסך הכול מעבר לשנים (N=24548)

| תחום/שנת בחינה            | 2008 | 2009 | 2010 | 2011 | סה"כ  |
|---------------------------|------|------|------|------|-------|
| N                         | 6181 | 6227 | 6030 | 6110 | 24548 |
| עברית-מתמטיקה             | 0.62 | 0.64 | 0.62 | 0.62 | 0.62  |
| עברית-אנגלית              | 0.53 | 0.56 | 0.51 | 0.55 | 0.53  |
| מתמטיקה-אנגלית            | 0.47 | 0.50 | 0.44 | 0.48 | 0.47  |
| ציון הבחינה הכללי-עברית   | 0.88 | 0.88 | 0.88 | 0.88 | 0.88  |
| ציון הבחינה הכללי-מתמטיקה | 0.87 | 0.89 | 0.87 | 0.87 | 0.87  |
| ציון הבחינה הכללי-אנגלית  | 0.72 | 0.74 | 0.7  | 0.73 | 0.72  |

ניתן לראות כי המתאמים בין חלקי הבחינה לבין הציון הסופי הם הגבוהים ביותר. מתאמים אלו מנופחים במידת מה, שהרי הציון הסופי מכיל בתוכו את ציוני החלקים השונים. אחר כך, המתאמים הגבוהים ביותר הם בין חלק העברית לחלק המתמטיקה ובין חלק העברית לחלק האנגלית. המתאמים הנמוכים ביותר (אך עדיין גבוהים) הם בין החלק המתמטיקה לחלק האנגלית.

## 2. דגימת הנוסחים

עד היום טרם נאספו נתונים על דגימת הנוסחים. מעתה ואילך נרצה לאסוף נתונים כאלה. כדוגמה ראשונית, בטבלה 4 מוצגים מספר הדגימות לפי נוסחים ויחס הדגימה המתוכנן במהלך רוב שנת הבחינות 2013 (מ-23.4.13 עד 26.12.2013). מועד זה חולק לשתי תקופות: תקופה א' בה הועבר גם נוסח B2 (מתחילת שנת הבחינות עד 24.10.13), ותקופה ב' בה הורד נוסח B2 מהמחזור (ר' ממצאי "שקילות נוסחים" להלן; מ-31.10.13 עד סוף התקופה). נערך מבחן חי בריבוע על מנת לבחון האם הדגימה בפועל עבדה כמתוכנן. שני המשתנים היו יחס בפועל לעומת יחס מתוכנן. שני המשתנים לא נמצאו קשורים באופן מובהק בתקופה הראשונה  $\chi^2(7, N=4452) = 8.94, p = .261$ , וכן בתקופה השנייה  $\chi^2(6, N=352) = 3.66, p = .721$ ; כלומר, הדגימה בפועל אינה שונה באופן מובהק מהדגימה המתוכננת; הדגימה עבדה כמתוכנן.

**טבלה 4: מספר דגימות לפי נוסחים ויחס הדגימה המתוכנן במהלך רוב שנת הבחינות 2013 בשתי תקופות**

### תקופה א'

| נוסח | הרכב | מספר נבחנים | יחס בפועל | יחס מתוכנן |
|------|------|-------------|-----------|------------|
| B2   | 1    | 265         | 0.06      | 0.06       |
| C2   | 1    | 265         | 0.06      | 0.06       |
| C2   | 2    | 273         | 0.06      | 0.06       |
| D2   | 1    | 277         | 0.06      | 0.06       |
| D2   | 2    | 287         | 0.06      | 0.06       |
| F1   | 1    | 1023        | 0.23      | 0.24       |
| G1   | 1    | 985         | 0.22      | 0.24       |
| E1   | 1    | 1077        | 0.24      | 0.24       |
| סה"כ |      | 4452        | 1         | 1          |

### תקופה ב'

| נוסח | הרכב | מספר נבחנים | יחס בפועל | יחס מתוכנן |
|------|------|-------------|-----------|------------|
| C2   | 1    | 17          | 0.05      | 0.06       |
| C2   | 2    | 28          | 0.08      | 0.06       |
| D2   | 1    | 21          | 0.06      | 0.06       |
| D2   | 2    | 22          | 0.06      | 0.06       |
| F1   | 1    | 83          | 0.24      | 0.25       |
| G1   | 1    | 95          | 0.27      | 0.25       |
| E1   | 1    | 86          | 0.24      | 0.25       |
| סה"כ |      | 352         | 1         | 1          |

### 3. סדר פרקים רנדומאלי

כמקרה בוחן, התמקדנו בנוסח D2 בשנת הבחינות 2011, שהועבר ל-1,340 נבחנים. בנוסח שמונה פרקים: חלק האנגלית (פרקים ENG1, ENG2), חלק העברית (HEB1, HEB2), חלק המתמטיקה (MATH1, MATH2) ושני פרקי פיילוט (PILOT1, PILOT2). כל אחד מ-1,340 הנבחנים קיבל את שמונת הפרקים המרכיבים את נוסח זה בסדר מקרי. אם אכן הסדר הוא מקרי לגמרי, אזי כל פרק אמור להופיע בכל אחד משמונת המיקומים האפשריים בבחינה בשכיחות של 12.5% (דהיינו, 167 פעם).

נערך מבחן חי בריבוע על מנת לבחון האם כל פרק הופיע בכול אחד משמונת המיקומים האפשריים בבחינה באופן רנדומאלי. שני המשתנים היו מספר הפרק (8 פרקים) ומיקום בבחינה (מקום 2 עד 9, כאשר מקום 1 לא נבחן שכן הוא תמיד מוקדש לפרק ההסברים). שני המשתנים לא נמצאו קשורים באופן מובהק  $\chi^2(49, N=10720) = 48.68, p = .486$ . ואכן, בטבלה 5 ניתן לראות פער קטן בין השכיחות בפועל לשכיחות המצופה.

**טבלה 5: שכיחות מספר פרק ומיקום הפרק בבחינה בפועל, לעומת המצופה במצב של היעדר קשר, בנוסח D2 בשנת בחינה 2011 (N=1,340)**

| מיקום הפרק |     |     |     |     |     |     |     | בפועל  |          |
|------------|-----|-----|-----|-----|-----|-----|-----|--------|----------|
| 9          | 8   | 7   | 6   | 5   | 4   | 3   | 2   |        |          |
| 158        | 158 | 165 | 162 | 169 | 188 | 158 | 182 | PILOT1 | מספר פרק |
| 186        | 180 | 161 | 156 | 157 | 158 | 174 | 168 | PILOT2 |          |
| 176        | 153 | 166 | 190 | 167 | 158 | 168 | 162 | ENG1   |          |
| 157        | 176 | 155 | 161 | 156 | 187 | 157 | 191 | ENG2   |          |
| 169        | 173 | 168 | 170 | 158 | 178 | 179 | 145 | HEB1   |          |
| 175        | 173 | 169 | 161 | 176 | 149 | 175 | 162 | HEB2   |          |
| 158        | 156 | 192 | 175 | 173 | 159 | 179 | 148 | MATH1  |          |
| 165        | 172 | 166 | 165 | 184 | 162 | 148 | 178 | MATH2  |          |
| 167        | 167 | 167 | 167 | 167 | 167 | 167 | 167 | מצופה  |          |

### 4. יציבות קושי הפריטים

על מנת לבחון האם קושי הפריטים השתנה מעבר לשנים, בדקנו את נוסח E1. נוסח E1 נבחר שכן הינו הנוסח היחיד שרץ באותה המתכונת יותר משלוש שנים (להבדיל מ-B1 ו-B2, C1 ו-C2, D1 ו-D2, שעברו הברכה<sup>1</sup> ולא רצו יותר משנתיים רצופות באותה המתכונת), דבר המאפשר לערוך השוואה של הקושי מעבר לזמן, בין שנת בחינה 2009 לבין שנת בחינה 2011. כמו כן, מדובר בנוסח המועבר עד היום. קושי פריט הוגדר כאחוז משיבים נכון. ככל שאחוז התשובות הנכונות נמוך יותר, פירוש הדבר שהפריט יותר קשה. בטבלה 6 מוצגים ממוצעים (וסטיות תקן) של אחוז התשובות הנכונות הממוצע בכל פרק ותחום בנוסח E1, בשנת בחינה 2009 לעומת שנת בחינה 2011.

<sup>1</sup> הברכה (doglegging) מהווה שינוי באופן חישוב הציון, אך לא בנוסח הבחינה עצמו. נוסח B2 הינו הברכה של C2, B1 של C1 ו-D2 של D1.

**טבלה 6: ממוצעים (וסטיות תקן) של אחוז התשובות הנכונות הממוצע על פי פרק, תחום ושנת בחינה בנוסח E1 (N=156)**

| שנת בחינה        |                  | N  | מספר פרק | תחום    |
|------------------|------------------|----|----------|---------|
| 2011             | 2009             |    |          |         |
| 58.54<br>(13.20) | 59.19<br>(13.20) | 30 | HEB3     | עברית   |
| 58.01<br>(12.76) | 59.67<br>(12.65) | 30 | HEB4     |         |
| 58.27<br>(12.87) | 59.43<br>(12.82) | 60 | סה"כ     |         |
| 58.97<br>(14.98) | 60.05<br>(15.17) | 23 | MATH3    |         |
| 57.53<br>(14.44) | 58.95<br>(14.18) | 23 | MATH4    | מתמטיקה |
| 58.25<br>(14.57) | 59.50<br>(14.53) | 46 | סה"כ     |         |
| 60.13<br>(12.42) | 61.91<br>(11.92) | 25 | ENG3     | אנגלית  |
| 58.37<br>(11.08) | 60.75<br>(10.04) | 25 | ENG4     |         |
| 59.25<br>(11.68) | 61.33<br>(10.93) | 50 | סה"כ     |         |

על מנת לבחון את ההבדל בין קושי הפריטים בין שתי התקופות (שנת בחינה 2009 לעומת שנת בחינה 2011) באופן יותר שיטתי, תוך בחינת הבדלים מובהקים סטטיסטית מעבר להצגה הגרפית התיאורית, ערכנו מבחני t מזווג (Matched Pairs). בטבלה 7 מוצגות תוצאות מבחני ה-t.

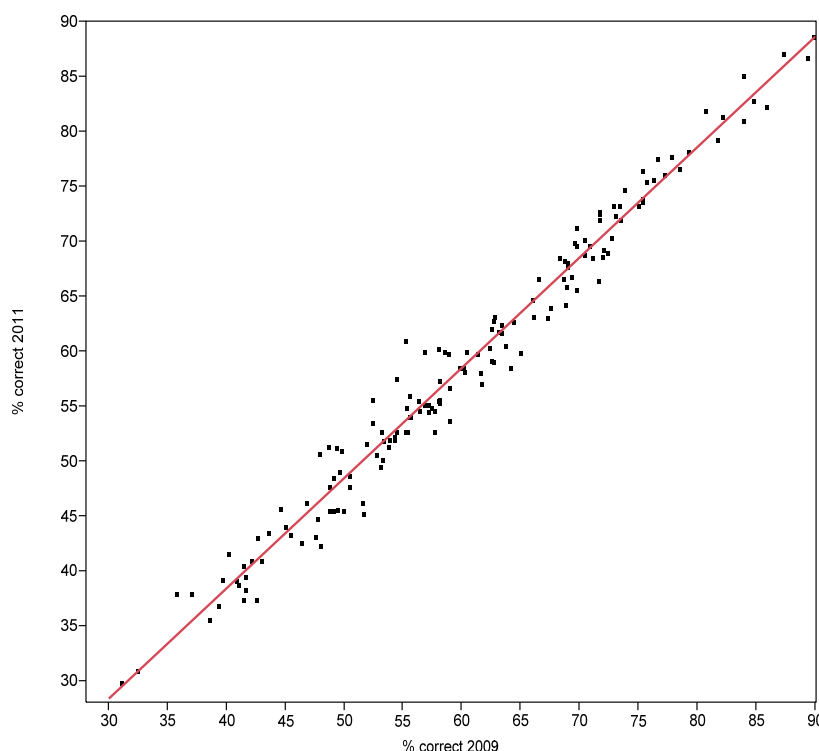
**טבלה 7: תוצאות מבחני ה-t לפי תחום – סטטיסטי המבחן, דרגות החופש, אלפא אפוסטריורית (P-value) והקשר בין אחוז התשובות הנכונות הממוצע בשתי שנות הבחינה**

| תחום    | t    | df | P-value     | r    |
|---------|------|----|-------------|------|
| עברית   | 4.20 | 59 | $p < 0.001$ | 0.99 |
| מתמטיקה | 4.07 | 45 | $p < 0.001$ | 0.99 |
| אנגלית  | 8.11 | 49 | $p < 0.001$ | 0.99 |

עברנו על הפריטים החריגים ביותר בכל התחומים, אך לא נמצא כל גורם שיטתי שיכול להסביר את חריגותם.

בלוח 5 מוצג לדוגמה אחוז העונים נכון בנוסח E1 בתחום האנגלית. ציר ה-X הינו אחוז העונים נכון בשנת בחינה 2009 וציר Y הינו אחוז העונים נכון בשנת בחינה 2011, כך שכל נקודה מסמלת פריט. ניתן לראות שבשנת בחינה 2011 אחוז העונים נכון נמוך מבשנת בחינה 2009.

**לוח 5: אחוז העונים נכון בכל אחד מ-50 הפריטים בתחום האנגלית בנוסח E1 בשנת בחינה 2009 ובשנת בחינה 2011 (N=4,644)**



**5. שקילות נוסחי הבחינה**

ניתוח שונות חד-גורמי ברמת מובהקות של 5% נערך על מנת לבחון אם קיימים הבדלים בין הממוצעים של ציוני הנבחנים בנוסחים השונים (מעבר לשנים, ממאי 2008 ועד מרץ 2012). המשתנה הבלתי תלוי, נוסח הבחינה, כלל 9 רמות: A1, B1, B2, C1, C2, D1, D2, E1 ו-F1. הניתוח נעשה בנפרד על כל משתנה תלוי: הציון בחלק העברית, הציון בחלק המתמטיקה, הציון בחלק האנגלית וציון הבחינה הכללי. טבלה 8 מתארת את תוצאות ניתוחי השונות.

**טבלה 8: תוצאות ניתוח השונות לפי תחום – סטטיסטי המבחן, אלפא אפוסטריורית (P-value) ואחוז השונות המוסברת**

| תחום              | F (8, 24539) | P-value      | אחוז השונות המוסברת |
|-------------------|--------------|--------------|---------------------|
| עברית             | 56.93        | $p < 0.0001$ | 1.82                |
| מתמטיקה           | 4.81         | $p < 0.0001$ | 0.16                |
| אנגלית            | 6.09         | $p < 0.0001$ | 0.20                |
| ציון הבחינה הכללי | 12.52        | $p < 0.0001$ | 0.41                |

בטבלה 9 מוצגות תוצאות הפוסט-הוק מסוג HSD (Honestly Significant Difference של Tukey) על ציון הבחינה הכללי. נוסחים שאין להם אות משותפת שונים במובהק זה מזה. מהטבלה עולה כי נוסח A1 שונה במובהק מכל הנוסחים האחרים. הפער בין הציון שחושב בפועל לנוסח A1 לבין ממוצע ציוני שאר הנוסחים עמד על כ-24 נקודות, בין רבע לחמישית סטיית תקן של הציונים. כאשר מאחדים את הנוסחים לפני ואחרי ההברכה (B1 ו-B2, C1 ו-C2, D1 ו-D2,

שהפכו לנוסח B, C ו-D, בהתאמה) כמוצג בטבלה 10, נראה כי בנוסף לנוסח A1, אף נוסח B המאוחד שונה במובהק מיתר הנוסחים מלבד F1. בטבלה 11, 12 ו-13 מופיע פוסט-הוק על ציון חלק העברית, המתמטיקה והאנגלית, בהתאמה.

**טבלה 9: פוסט-הוק מסוג HSD על ציון הבחינה הכללי**

| ציון ממוצע |   |   |   | נוסח |
|------------|---|---|---|------|
| 414.56     |   |   | A | D2   |
| 410.17     | C | B | A | C2   |
| 409.67     | C |   | A | D1   |
| 408.80     | C | B | A | F1   |
| 408.08     | C |   | A | E1   |
| 405.28     | C | B | A | C1   |
| 401.03     |   | B |   | B1   |
| 400.43     | C | B |   | B2   |
| 385.61     | D |   |   | A1   |

**טבלה 10: פוסט-הוק מסוג HSD על ציון הבחינה הכללי לאחר איחוד נוסחים**

| ציון ממוצע |   |   |   | נוסח |
|------------|---|---|---|------|
| 411.66     |   |   | A | D    |
| 408.80     |   | B | A | F1   |
| 408.68     |   |   | A | C    |
| 408.08     |   |   | A | E1   |
| 400.85     |   | B |   | B    |
| 385.61     | C |   |   | A1   |

**טבלה 11: פוסט-הוק מסוג HSD על ציון חלק העברית**

| ציון ממוצע |   |   |   | נוסח |
|------------|---|---|---|------|
| 83.33      |   |   | A | D2   |
| 82.51      |   |   | B | F1   |
| 82.35      |   |   | B | D1   |
| 81.17      |   |   | B | E1   |
| 80.79      |   | C | B | C2   |
| 78.37      | D | C |   | C1   |
| 77.70      | D |   |   | B2   |
| 77.60      | D |   |   | B1   |
| 73.28      | E |   |   | A1   |

**טבלה 12: פוסט-הוק מסוג HSD על ציון החלק המתמטיקה**

| ציון ממוצע | נוסח |   |    |
|------------|------|---|----|
| 83.70      | B    | A | C1 |
| 83.52      |      | A | C2 |
| 83.08      |      | A | B1 |
| 82.67      | C    | B | D2 |
| 82.67      | C    | B | B2 |
| 81.86      | C    | B | A1 |
| 81.61      | C    | B | D1 |
| 81.41      | C    | B | E1 |
| 80.62      | C    |   | F1 |

**טבלה 13: פוסט-הוק מסוג HSD על ציון חלק האנגלית**

| ציון ממוצע | נוסח |   |    |
|------------|------|---|----|
| 83.51      |      | A | E1 |
| 83.05      | B    | A | F1 |
| 82.09      | C    | B | C1 |
| 82.04      |      | B | D2 |
| 81.96      |      | B | D1 |
| 81.74      | C    | B | C2 |
| 81.55      | C    | B | B2 |
| 81.34      | C    | B | B1 |
| 79.62      | C    |   | A1 |

השונוות הרבה ביותר בין הנוסחים ניכרה בחלק העברית, כאשר בחלק המתמטיקה ובחלק האנגלית ניכר יותר דמיון בין הנוסחים. החריגות של נוסח A1 וכן של נוסח B1 ו-B2 ניכרה בעיקר בחלק העברית.

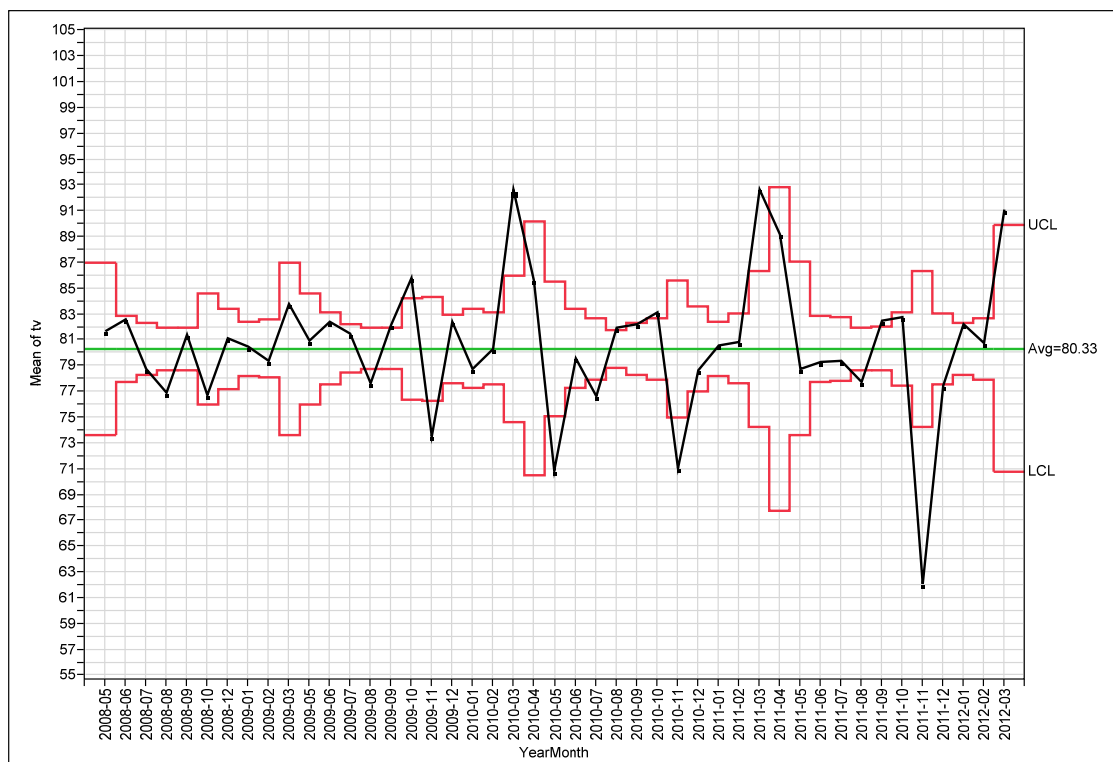
## 6. יציבות ציונים לאורך זמן

### א. טבלאות בקרה מסוג שוהארט על הציונים

להלן ארבעה לוחות ובהם טבלאות בקרה מסוג שוהארט המציגים את ממוצעי ארבעת התחומים (כל תחום בגרף נפרד: עברית, מתמטיקה, אנגלית וציון הבחינה הכללי, לוחות 6-9, בהתאמה) על פני ציר הזמן (לפי חודשים, החל ממאי 2008 ועד מרץ 2012). כאקספלורציה ראשונית של נתונינו, בחרנו להתחיל עם טבלת בקרה בסיסית המכילה את כל השנים ואת כל הנוסחים.

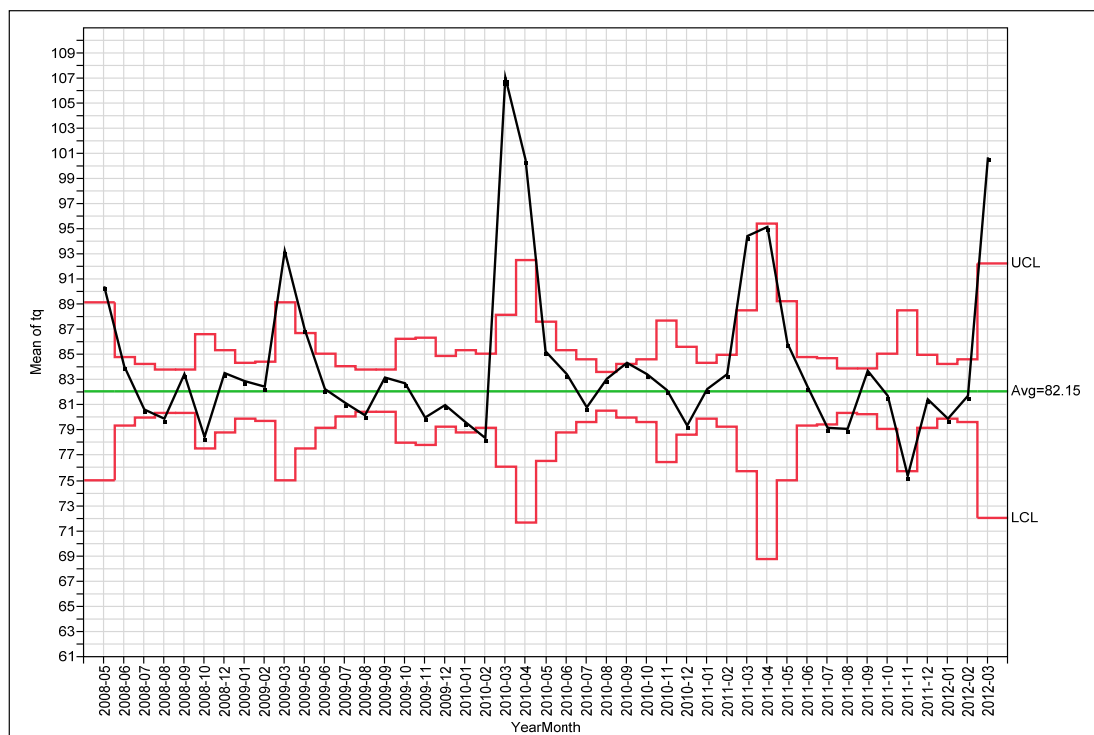
כפי שראינו לעיל בטבלה 2, הממוצעים וסטיות התקן בכל שנת בחינה בנפרד ומעבר לשנות בחינה הם דומים מאוד. קו הבסיס (Baseline) הינו ממוצע ציוני המבחן מעבר לארבעת השנים (הקו הירוק; הערך המצופה). כל נקודה מייצגת את ממוצע ציוני הנבחנים בתחום האמור באותו החודש (הנקודות מחוברות זו לזו באמצעות הקו השחור). סטיית התקן מעבר לחודשים וגודלי המדגם בכל חודש שימשו לחישוב הגבולות. הגבולות הינם 3 טעויות תקן מעל (Upper Control Limit – UCL) ומתחת (Lower Control Limit – LCL) לקו הבסיס (סטיות התקן לאחר תיקון לפי גודל המדגם – חלוקה בשורש N). להסבר מפורט יותר על אופן חישוב הגבולות, ר'נספח 3.

**לוח 6: טבלת בקרה עבור ציון חלק העברית (N=24548)**



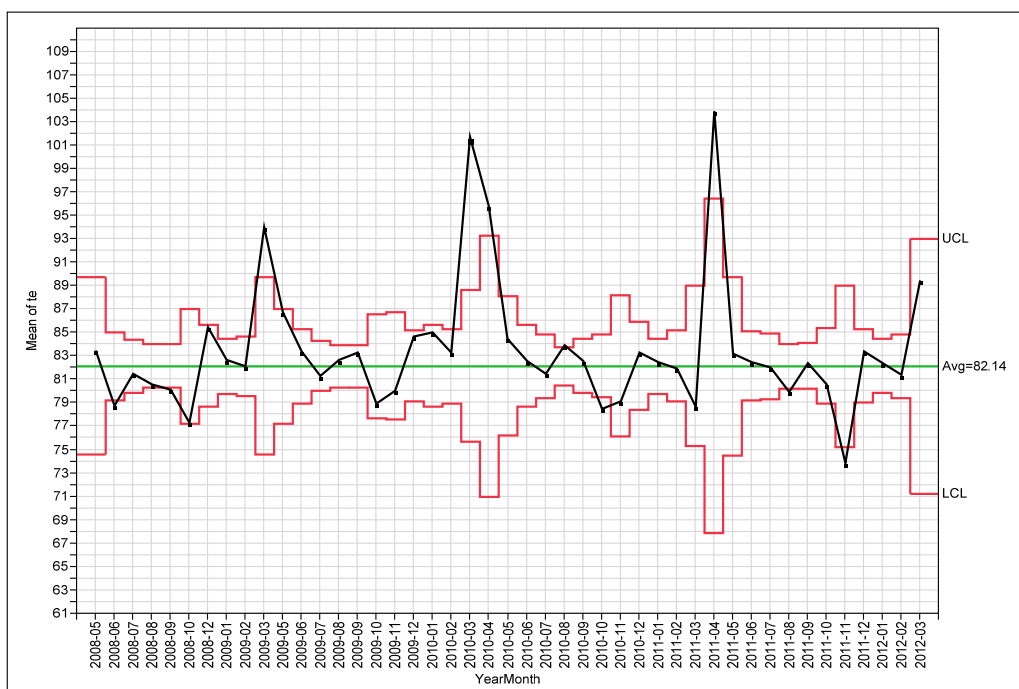
כפי שניתן לראות בגרף, לחלק העברית היו 17 מקרים של "יציאה משליטה" (Out of Control) מתוך 45 (37.8%), כאשר 8 מתוכם חרגו מעל לגבול העליון ו-9 מתוכם מתחת לגבול התחתון. לפירוט החריגות, ר' נספח 4.

**לוח 7: טבלת בקרה עבור ציון חלק המתמטיקה (N=24548)**



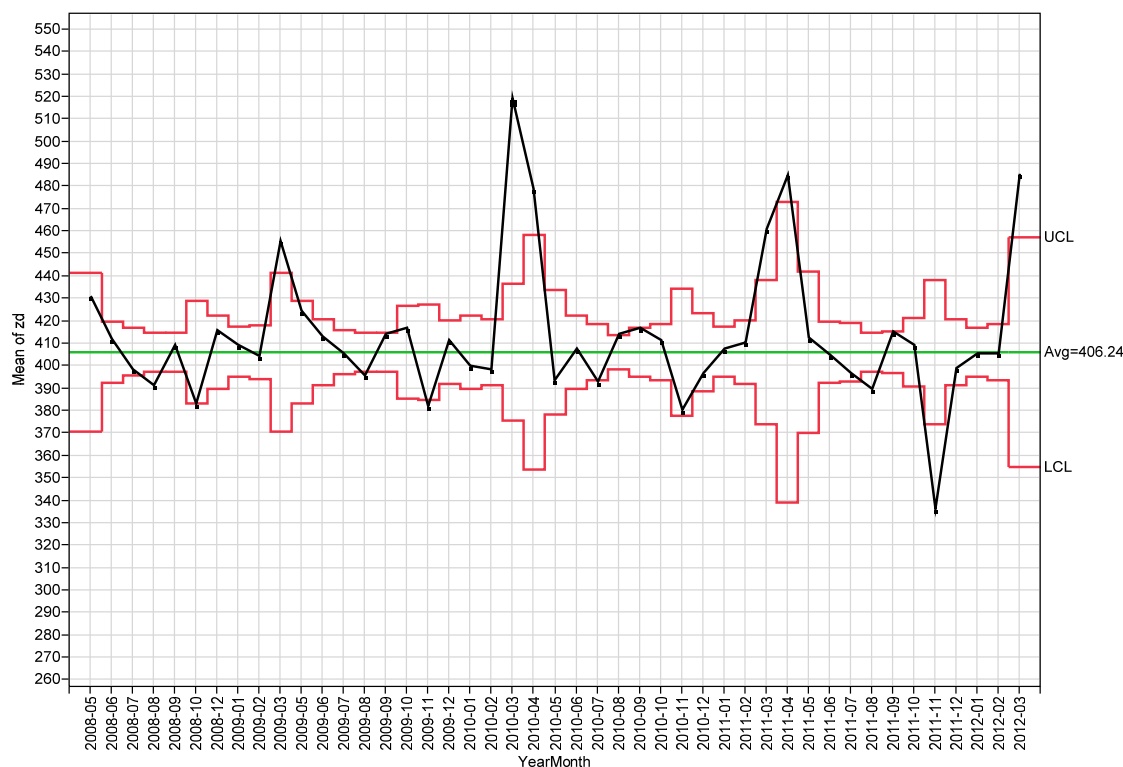
כפי שניתן לראות בגרף, לחלק המתמטיקה היו 14 מקרים של "יציאה משליטה" מתוך 45 (31.1%), כאשר 8 מתוכם חרגו מעל לגבול העליון ו-6 מתוכם מתחת לגבול התחתון. לפירוט החריגות, ר' נספח 4.

**לוח 8: טבלת בקרה עבור ציון חלק האנגלית (N=24548)**



כפי שניתן לראות בגרף, לחלק האנגלית היו 10 מקרים של "יציאה משליטה" מתוך 45 (22.2%), כאשר 5 מתוכם חרגו מעל לגבול העליון ו-5 מתוכם מתחת לגבול התחתון. לפירוט החריגות, ר' נספח 4.

**לוח 9: טבלת בקרה עבור ציון הבחינה הכללי (N=24548)**



לציון הבחינה הכללי היו 12 מקרים של "יציאה משליטה" מתוך 45 (26.7%), כאשר 7 מתוכם חרגו מעל לגבול העליון ו-5 מתוכם מתחת לגבול התחתון. לפירוט החריגות, ר' נספח 4.

להסתכלות מפורטת יותר שרטטנו גם טבלאות בקרה לכל נוסח בנפרד, טבלאות הבקרה המלאות בנספח 5.

## ב. רגרסיה ליניארית על ציון הבחינה הכללי

נערך ניתוח רגרסיה מרובה על מנת לנבא את ציון הבחינה הכללי. המנבאים היו מגדר, הארכת זמן ומוסד אקדמי בו נבחנו המשתתפים. משוואת הרגרסיה נמצאה מובהקת,  $R^2 = 0.17$ ,  $\text{adjusted } R^2 = 0.17$ ,  $F(40, 18710) = 84.46$ ,  $p < 0.001$ . בטבלה 14 מובהקות מקדמי המתאם של המנבאים באמצעות ניתוח שונות.

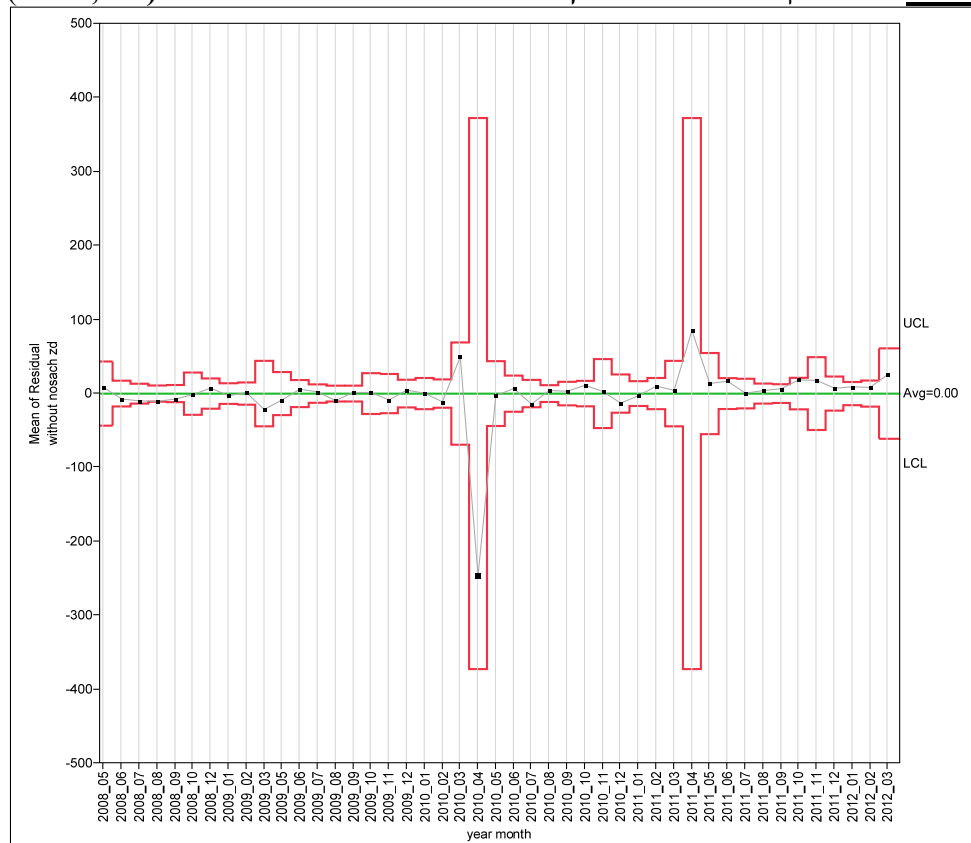
| <b>טבלה 14: מובהקות המקדמים ברגרסיה הליניארית לניבוי ציון הבחינה הכללי (N=18750)</b> |           |          |            |
|--------------------------------------------------------------------------------------|-----------|----------|------------|
| <b>משתנה</b>                                                                         | <b>df</b> | <b>F</b> | <b>Sig</b> |
| <b>מגדר</b>                                                                          | 1, 18748  | 1081.06  | <.0001     |
| <b>הארכת זמן</b>                                                                     | 2, 18747  | 123.87   | <.0001     |
| <b>מוסד אקדמי בו נבחנו המשתתפים</b>                                                  | 37, 18712 | 46.85    | <.0001     |

בנספח 6 מוצג הפירוט המלא אודות מקדמי המתאם של תתי המנבאים, טעות התקן שלהם, מובהקותם באמצעות מבחן  $t$  ומקדמיהם המתוקננים, לרבות קו הבסיס. למרות ריבוי תתי המקדמים המובהקים, ערכי המקדמים המתוקננים שהתקבלו היו קטנים. המשתנים שנמצאו כמשפיעים ביותר כפי שבא לידי ביטוי בערכי ביתא גבוהים, הינם אישה ( $\beta = -0.23$ ) קמפוס OJ ( $\beta = 0.17$ ), קמפוס CL ( $\beta = 0.15$ ), קמפוס MO ( $\beta = -0.14$ ), קמפוס IT ( $\beta = 0.13$ ), ללא תוספת זמן ( $\beta = 0.11$ ) וקמפוס TO ( $\beta = -0.10$ ).

### ג. טבלת בקרה על השאריות של ציון הבחינה הכללי מהגרסיה הליניארית

בלוח 10 מוצגת טבלת בקרה מסוג שוהארט על השאריות (Residuals) של ציון הבחינה הכללי על פני ציר הזמן (לפי חודשים, החל ממאי 2008 ועד מרץ 2012) (Gutentag, 2013). חישוב השאריות נעשה כדלקמן: ההפרש בין הציון המנובא של כל נבחן על פי הרגרסיה הליניארית דלעיל לבין הציון הנצפה (הציון שהנבחן קיבל באמת). קו הבסיס (הערך המצופה) הוא ממוצע השאריות מעבר לחודשים, והוא כמובן שווה ל-0 מעצם הגדרת מודל הרגרסיה. כל נקודה מייצגת את ממוצע השאריות באותו החודש. הגבולות הינם 3 טעויות תקן מעל (Upper Control Limit – UCL) ומתחת (Lower – LCL Control Limit) לקו הבסיס (סטיות התקן לאחר תיקון לפי גודל המדגם – חלוקה בשורש N). להסבר מפורט יותר על אופן חישוב הגבולות, ר' נספח 3.

**לוח 10: טבלת בקרה על שאריות ציון הבחינה הכללי ברגרסיה הליניארית (N=18,751)**



מעיון בלוח ניכר כי אף נקודה אינה חורגת מחוץ לגבולות הבקרה (מיעוטן על הגבול ממש).

## 7. הישגי נבחנים חוזרים

בבחינות שבין 2008 (אפריל) ל-2012 (מרץ), היו במימ"ד 24,548 היבחנות. 23,949 מתוכם היו בחינות ראשונה. יתר העברות המבחן, 2.4% מסך הבחינות (N=599), היו בחינות חוזרות, כלומר, הועברו לנבחנים שכבר נבחנו בעבר בבחינת המימ"ד במהלך שנים אלה. בטבלה 15 מוצגת שכיחות מספר הבחינות של נבחני מימ"ד בשנות בחינה 2008-2012.

**טבלה 15: שכיחות מספר הבחינות של נבחני מימ"ד בשנות בחינה 2008-2012**

| מספר הבחינות | שכיחות | אחוז  |
|--------------|--------|-------|
| 1            | 23,949 | 97.56 |
| 2            | 554    | 2.26  |
| 3            | 43     | 0.18  |
| 4            | 2      | 0.01  |
| סה"כ         | 24,548 | 100   |

בטבלה 16 מוצגים ממוצע, סטיית תקן, מינימום, חציון, מקסימום ומספר נבחנים לפי מספר העברה (בחינה ראשונה, שנייה, שלישית ורביעית) ותחום (אנגלית, מתמטיקה, עברית וחציון הבחינה הכללי). בכל התחומים ניכרת מגמה של ירידה בציונים ככל שמספר החזרות על הבחינה עולה. חשוב לזכור, כי הנתונים המוצגים כאן הם נתונים בין-נבחניים ולא בתוך אותה קבוצת נבחנים (תוך-נבחניים); על כן, מגמת ירידה זו אינה מצביעה כי הנבחנים החוזרים יורדים בהישגיהם מבחינה לבחינה, אלא משקפת את העובדה שרמת ההישגים של הנבחנים הבוחרים לגשת לבחינה חוזרת נמוכה יחסית. זאת באופן טבעי, כיוון שנבחנים שהשיגו בבחינה ציון גבוה מספיק לצורכיהם אינם צריכים לשפר את ציונם ולא יטו לגשת ולהיבחן בשנית.

**טבלה 16: ממוצע, סטיית תקן, מינימום, חציון, מקסימום ומספר נבחנים לפי מספר העברה ותחום בחינה 2008-2012**

| מספר העברה   | תחום    | ממוצע  | סטיית תקן | מינימום | חציון | מקסימום | N      |
|--------------|---------|--------|-----------|---------|-------|---------|--------|
| בחינה ראשונה | אנגלית  | 80.53  | 19.87     | 21      | 80    | 125     | 23,949 |
|              | מתמטיקה | 82.32  | 21.27     | 19      | 81    | 126     |        |
|              | עברית   | 82.33  | 23.27     | 25      | 80    | 125     |        |
|              | כללי    | 407.38 | 106.90    | 50      | 404   | 650     |        |
| בחינה שנייה  | אנגלית  | 73.13  | 18.27     | 25      | 73    | 125     | 554    |
|              | מתמטיקה | 75.47  | 18.58     | 38      | 75    | 125     |        |
|              | עברית   | 75.22  | 18.52     | 41      | 72    | 125     |        |
|              | כללי    | 364.67 | 90.28     | 114     | 361   | 632     |        |
| בחינה שלישית | אנגלית  | 61.95  | 15.65     | 38      | 56    | 101     | 43     |
|              | מתמטיקה | 69.00  | 19.52     | 42      | 67    | 125     |        |
|              | עברית   | 67.40  | 16.50     | 47      | 61    | 108     |        |
|              | כללי    | 312.91 | 78.86     | 172     | 310   | 503     |        |
| בחינה רביעית | אנגלית  | 55.50  | 10.61     | 48      | 55.5  | 63      | 2      |
|              | מתמטיקה | 86.50  | 7.78      | 81      | 86.5  | 92      |        |
|              | עברית   | 69.00  | 21.21     | 54      | 69    | 84      |        |
|              | כללי    | 341.50 | 19.09     | 328     | 341.5 | 355     |        |

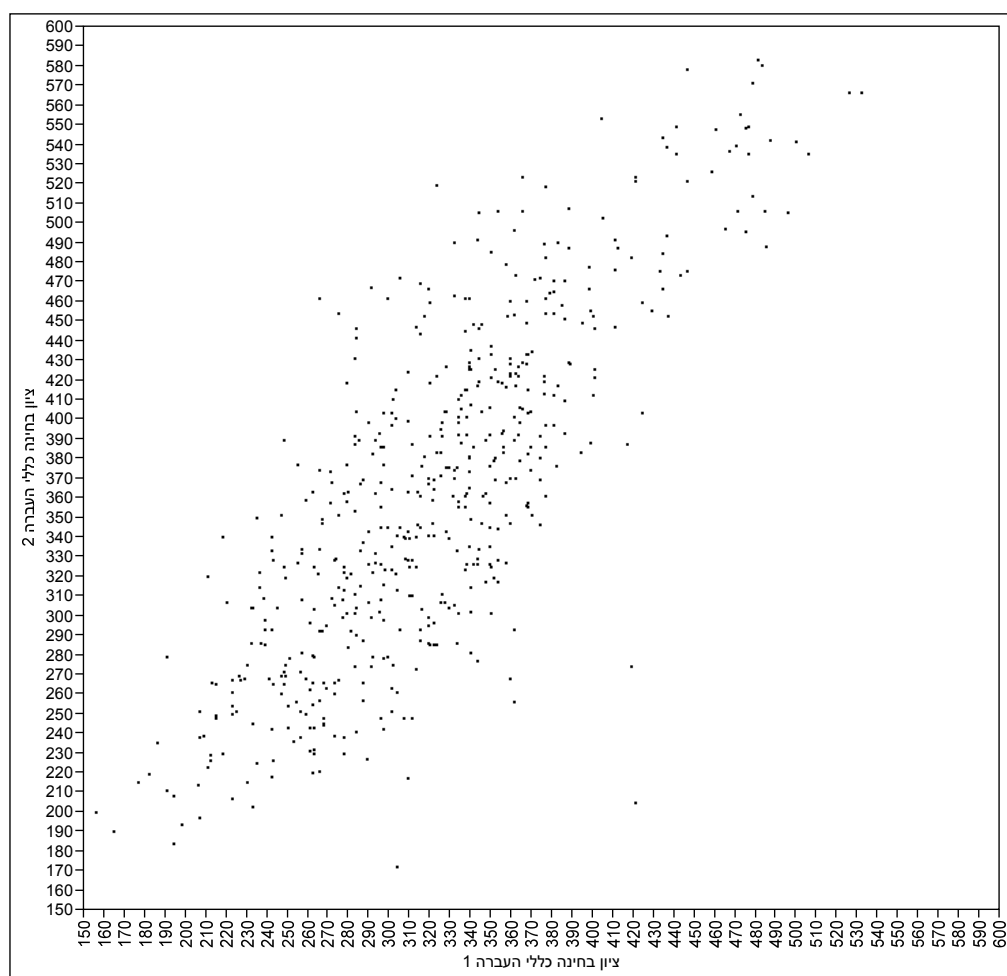
על מנת לעמוד על מידת השיפור בהישגים מבחינה לבחינה, יש לערוך בדיקה בתוך קבוצת הנבחנים החוזרים בלבד. בטבלה 17 מוצגים הפרשי ציון הבחינה הכללי של הנבחנים החוזרים בין העברות. מאחר ששני נבחנים בלבד ניגשו לבחינה 4 פעמים בתקופה האמורה לא הושוו נתוני ההעברה הרביעית עם יתר ההעברות.

**טבלה 17: ממוצע, סטיית תקן, חציון, מתאם פירסון ומספר נבחנים של הפרשי ציון הבחינה הכללי בין העברות בשנות בחינה 2008-2012**

| הפרשי ציון הבחינה הכללי | ממוצע | סטיית תקן | חציון | מתאם פירסון | N   |
|-------------------------|-------|-----------|-------|-------------|-----|
| בין העברה 2 להעברה 1    | 41.93 | 56.05     | 39    | 0.78        | 554 |
| בין העברה 3 להעברה 1    | 29.86 | 52.52     | 27    | 0.75        | 43  |
| בין העברה 3 להעברה 2    | 13.88 | 50.02     | 13    | 0.77        |     |

לוח 11 מציג בקרב הנבחנים החוזרים את הקשר בין ציון הבחינה הכללי בהעברה הראשונה לבין ציון הבחינה הכללי בהעברה השנייה.

**לוח 11: הקשר בין ציון הבחינה הכללי בבחינה הראשונה לבין הבחינה השנייה בקרב הנבחנים החוזרים בשנות בחינה 2008-2012 (N=553)**



בטבלה 18 מוצגים הפרשי זמני ההעברות בימים של הנבחנים החוזרים. מאחר ששני נבחנים בלבד לקחו את הבחינה 4 פעמים, לא הושוו נתוני ההעברה הרביעית עם יתר ההעברות.

**טבלה 18: חציון, ממוצע, סטיית תקן ומספר נבחרים של הפרשי זמני ההעברות בימים בשנות בחינה 2008-2012**

| הפרשי ימים           | חציון | ממוצע | סטיית תקן | N   |
|----------------------|-------|-------|-----------|-----|
| בין העברה 2 להעברה 1 | 329   | 395   | 286       | 553 |
| בין העברה 3 להעברה 1 | 341   | 475   | 227       | 43  |
| בין העברה 3 להעברה 2 | 132   | 182   | 111       |     |

**מעקב אחר נוסחי הבחינה שנבחרים חוזרים מקבלים**

49 נבחרים (0.2% מכלל הנבחרים, 8.9% מהנבחרים החוזרים) קיבלו את אותו נוסח פעמיים, כאשר 41 נבחרים קיבלו אותו נוסח בהעברה הראשונה והשנייה, 6 קיבלו אותו נוסח בהעברה הראשונה והשלישית, 1 קיבל את אותו נוסח בהעברה הראשונה והרביעית ו-1 קיבל את אותו נוסח בהעברה השנייה והרביעית. ממוצע משך הזמן בימים בין העברת הנוסחים הוזהים הינו 360.22 (סטיית תקן 146.15), חציון 346.00. משך הזמן המינימאלי בין שתי ההעברות 16 יום והמקסימאלי 820 יום, כאשר 20 מהנבחרים שקיבלו נוסחים זהים נבחנו באותה הבחינה פחות משנה לאחר שנבחנו בה לראשונה.

## דיון

הדיון בעבודה זו מחולק בהתאם לפרקי התוצאות. בכל סעיף יש התייחסות לשגרת הבדיקה המומלצת כדי לבצע את בקרת האיכות המתאימה באופן שוטף. בשגרת העבודה, אם מתגלה נתון חריג תתקבל התראה ("הקפצת דגל") הדורשת התייחסות ובדיקה ממוקדת לסיבת הבעיה.

### סטטיסטיקה תיאורית

הנוסחים שכרגע בשימוש שוטף הינם B2, C2 ו-D2 (החלו באותה תקופה – מאי 2010), E1 (הוותיק ביותר – מאי 2009) ו-F1 (החדש ביותר – מרץ 2011). אין הבדלים משמעותיים בין השנים בסטטיסטיים – ממוצע וסטיית תקן – של הציגונים המחושבים. יש חודשים בהם בשל מיעוט הנבחנים בשנות הבחינה השונות – כגון בחודש מרץ, אפריל ומאי (מספר הנבחנים בכל אחד מהחודשים נמוך מ-200) ובחודש נובמבר (מספר הנבחנים נמוך מ-250) – ישנו קושי בבקרה: הממוצע החודשי סוטה מהממוצע הכללי, וסטיית התקן תנודתית יותר. באשר למתאמים, נמצאו מתאמים בינוניים-גדולים בין כל חלקי המבחן: עברית, מתמטיקה ואנגלית, ובינם לבין ציון הבחינה הכללי.

### דגימת נוסחים

לצורך המחקר אספנו נתונים על דגימת הנוסחים אשר לא נאספו קודם לכן. מעתה ואילך נרצה לאסוף נתונים כאלה באופן שגרתי.

שגרת עבודה: תתבצע בדיקה תקופתית. תהא "הקפצת דגל" אם נוסח מסוים לא נדגם בכלל, או במצב של דגימת יתר. בנוסף, תהיה התראה האם ההתפלגות בפועל של הנוסחים המועברים, כלומר "ההתפלגות המצויה", שונה במובהק מההתפלגות הרצויה (כפי שהוגדרה מראש) באמצעות מבחן חי בריבוע.

### סדר הפרקים

לצורך המחקר הנוכחי, התמקדנו בנוסח D2 בשנת 2011. לא נמצא קשר בין מספר הפרק ומיקומו בבחינה, כך שכל פרק הופיע בכול המקומות האפשריים בבחינה. שגרת עבודה: תתבצע בדיקה תקופתית. תהא "הקפצת דגל" אם פרק מסוים מופיע מתחת/מעל לסף מסוים.

### קושי הפריטים

לצורך המחקר הנוכחי, התמקדנו בנוסח E1 בשתי שנות בחינה שונות (שנת בחינה 2009 לעומת שנת בחינה 2011). נמצא שבשנת בחינה 2011 אחוז התשובות הנכונות הממוצע היה נמוך באופן מובהק בכאחוז משנת בחינה 2009. נראה כי אם כבר חלה תמורה מעבר לשנים, הרי שהנוסח שנבדק הפך לקשה יותר ולא לקל יותר. ממצא זה אף יכול להוות עדות נגד לדליפת פריטים. יתכן וירידה בהצלחה בבחינה נובעת משינויים באוכלוסיה הנבחנת לאורך השנים. בדיקות שגרה: בדיקה תקופתית. תהא "הקפצת דגל" אם נמצא פריט החורג באופן דרמטי בהשוואה לשנות בחינה קודמות ברמת הקושי.

## שקילות הנוסחים

לכל חלק בחינה בנפרד (עברית, מתמטיקה, אנגלית וציון הבחינה הכללי), נמצא כי נוסחים שונים נבדלים אלו מאלו באופן מובהק בציונם הממוצע, אך גודל האפקט קטן. השונות הרבה ביותר בין הנוסחים ניכרה בחלק העברית, כאשר בחלק המתמטיקה ובחלק האנגלית היה דמיון רב יותר בין הנוסחים. מבחני המשך איששו החלטה קודמת להוריד את נוסח A1 ממחזור ההעברות כליל. כמו כן, בעקבות ניתוח זה נוסח B2 הורד ותוקן חלק העברית בו.

בדיקות שגרה: בחינה תקופתית של ההבדלים בין הנוסחים השונים. תהא "הקפצת דגל" באם נוסח ימצא שונה מיתר הנוסחים באחד מהתחומים או בציון הבחינה הכללי.

## יציבות ציונים מעבר לזמן

**א. טבלאות בקרה מסוג שוהארט על הציון הגולמי.** בחודשים מרץ, אוגוסט וספטמבר היו הכי הרבה חריגות, ואילו בחודשים פברואר, יוני ודצמבר היו הכי מעט חריגות. אחוז החריגות נע בין 22-38%, כך שהחריגות הרבות ביותר נמצאו בחלק העברית, אחר כך בחלק המתמטיקה, אחר כך בציון הבחינה הכללי ולבסוף בחלק האנגלית. בחלק העברית היו יותר חריגות מתחת לגבול התחתון, בחלק המתמטיקה ובציון הכללי היו יותר חריגות מעל לגבול העליון, ואילו בחלק האנגלית מספר החריגות מעל ומתחת לגבול התחתון היה דומה.

בדיקות שגרה: ליצור מנגנון ("הקפצת דגל") שיתריע באופן תקופתי כאשר התצפיות בחודש/העברה מסוימים יוצאות מגבולות הבקרה.

**ב. ניבוי ציון הבחינה הכללי באמצעות רגרסיה ליניארית.** נמצא כי שלושה חזאים: מגדר, תוספת זמן ומוסד אקדמי בו התקיימה הבחינה (הנבחנים לרוב נבחנים במוסד בו הם מעוניינים ללמוד) מסבירים כמעט 20% מהשונות במשתנה ציון הבחינה הכללי בבחינת המימ"ד. רמות המשתנים שניבאו במידה הרבה ביותר הינם אישה, ללא תוספת זמן ומוסדות אקדמיים ייחודיים לאוכלוסיות ספציפיות (המגזר החרדי, המגזר הערבי ואוכלוסיה מוטית טכנולוגיה).

בדיקות שגרה: בחינת אחוז השונות המוסברת על ידי משוואת הרגרסיה, כך שחריגה משמעותית מאחוז השונות המוסברת שנמצא במחקר הנוכחי "תקפיץ דגל".

**ג. טבלאות בקרה מסוג שוהארט עבור שאריות ציון הבחינה הכללי.** כעת, בחנו האם השונות שלא הוסברה באמצעות מגדר, תוספת זמן ומוסד אקדמי היא "בשליטה" (In Control). עבור כל נבדק חושב ההפרש בין הציון האמיתי בבחינת המימ"ד לבין הציון שנבא לו באמצעות הרגרסיה הליניארית מסעיף ב'. משרטוט טבלת בקרה מסוג שוהארט על השאריות עלה כי עתה היו מעט מאוד חריגות מעל ומתחת לגבולות הבקרה. כך, שונות משתנה ציון הבחינה הכללי הוסברה באמצעות שלושת החזאים – מגדר, תוספת זמן ומוסד אקדמי, ואילו השונות שלא הוסברה על ידם (השונות הבלתי מוסברת) נמצאה "בשליטה".

בדיקות שגרה: ליצור מנגנון ("הקפצת דגל") שיתריע באופן תקופתי כאשר השאריות בחודש/העברה מסוימים יוצאות מגבולות הבקרה.

## **נבחנים חוזרים**

מיעוט הבחינות (2%, N=599) היו בחינות חוזרות (ככל הנראה מאחר שהבחינה היא "בסיכון נמוך"). באופן טבעי, רמת ההישגים של הנבחנים הבוחרים לגשת לבחינה חוזרת נמוכה יחסית. המתאם מבחן-מבחן חוזר בציון הבחינה הכללי בקרב נבחנים חוזרים נמצא גבוה. 9% מהנבחנים החוזרים (49 נבחנים) קיבלו את אותו נוסח פעמים, כאשר ההפרש הממוצע בין ההעברות היה של כשנה בממוצע (20 מתוכם קיבלו את אותו נוסח בהפרש של פחות משנה).

**בדיקות שגרה:** שיפור הנבחנים החוזרים יבחן באופן שגרתי; ליצור מנגנון שימנע מנבחן לקבל את אותו הנוסח, או שיתריע בדעיכה ("הקפצת דגל") כי נבחן קיבל את אותו הנוסח.

## **משתנים מסבירים ומשתנים מבוקרים**

את הסיבות לנתונים חריגים המתגלים בבקרה ניתן לחלק לשני סוגים. הסוג הראשון הינו משתנים מסבירים, המאפשרים להסביר נתונים הנראים חריגים כפועל יוצא מטבע האוכלוסייה המסוימת שעשתה את המבחן. הסוג השני הוא משתנים מבוקרים, המאפשרים להסביר נתונים הנראים חריגים כנובעים מנוסח המבחן או מבעיה טכנית או אחרת בכיול הציונים ובחישובם.

### **משתנים מסבירים**

#### **1. עונתיות:**

א. **חודש העברה.** ישנם חודשי בחינה פופולאריים יותר (למשל, אוגוסט וספטמבר) וישנם חודשים פופולאריים פחות (למשל, מרץ ואפריל). בהתאמה, בזמני הבחינה המבוקשים יהיו יותר נבחנים מאשר בזמנים המבוקשים פחות (ר' גודל המדגם להלן). כמו כן, יתכן והנבחנים בחודשים השונים נבדלים זה מזה במוטיבציה שלהם להיבחן בהתאם לתאריך בו הם מתכוונים להירשם ללימודים.

ב. **גודל המדגם.** יש עונות בהם יש יותר/פחות נבחנים, בעיקר כתלות במועדי ההרשמה במוסדות האקדמיים. בשל סיבות סטטיסטיות (חוק המספרים הגדולים), נצפה לתנודות גבוהות יותר בציון הממוצע של הנבחנים ככל שמספר הנבחנים קטן יותר, ולהפך. גודל המדגם משפיע על מידת ייצוגיות המדגם.

2. **הבדלים בין אישיים.** בני אדם נבדלים אלו מאלו ביכולותיהם. תוצאות המבחנים מבטאות מגוון ביצועים זה.

3. **מוסד לימודים.** מבחן המימ"ד הועבר בתקופה הנבחנת ב-38 מוסדות אקדמיים בישראל הפרושים בכל רחבי הארץ, המלמדים תחומים שונים ואשר פונים לאוכלוסיות שונות. מאחר ונבחנים נוטים להיבחן במוסד בו ירצו ללמוד בעתיד, עושר המוסדות מרמז על עושר ההון האנושי בהם (ר' לעיל הבדלים בין אישיים). למשל, במדגם יש מוסדות טכנולוגיים גבוהים לצד מוסדות גבוהים לחינוך; מוסד גבוה למגזר הערבי לצד מוסד גבוה הפונה לאוכלוסייה יהודית-חרדית ועוד.

4. **תחום הבחינה.** הבחינה מודדת שלושה תחומים: עברית, מתמטיקה ואנגלית. ישנו מתאם בין שלושת תחומי הבחינה. מחד, נבחנים שונים נבדלים במידה בה הם טובים בכול תחום ותחום (ר' לעיל הבדלים בין אישיים); ומאידך, מעבר לנבחנים ניתן לראות מגמה לפיה הציון הממוצע בעברית נמוך מאשר הציון הממוצע באנגלית ובמתמטיקה. כמו כן, בעוד שבעברית ובמתמטיקה ההתפלגות היא נורמאלית בקירוב, באנגלית ההתפלגות הינה דו שיאית.

5. **נוסח הבחינה.** למרות שיש שאיפה שנוסחי הבחינה יהיו שקולים זה לזה, ואף לפני ההשקה מכילים ואומדים את הבחינות זו ביחס לזו, לעיתים נוסחי הבחינה אינם שקולים. למשל, נוסח A1 נתגלה עם הזמן כנוסח שמניב ציון ממוצע נמוך במובהק מאשר נוסחי הבחינה האחרים, ולכן הוסר מהמחזור כליל. כמו כן, נמצא כי נוסח B2 מפגין ציון ממוצע נמוך במובהק מנוסחי הבחינה האחרים בחלק העברית, ולכן הוסר מהמחזור לתיקון.
6. **בעיה טכנית.** כל הבדל נוסף שאינו יכול להיות מוסבר על ידי אחד מהגורמים הקודמים מעלה את ההשערה שיתכן ומקורו בבעיה טכנית כזו או אחרת. בעיה טכנית יכולה להיגרם בשל טעות מכונה או בשל טעות אנוש.

### שגרת בקרות אוטומטית

- השלב הבא הינו בניית מערכת שתבקר את התהליך באופן אוטומטי. המערכת תפעל בשלושה שלבים:
- (א) **שליפת הנתונים ממערכת המידע לשרת.** הפקת נתונים מבסיס הנתונים באופן אוטומטי אחת לתקופה.
- (ב) **עיבוד הנתונים בשרת.** עיבוד הנתונים בכל תקופה על בסיס הבקרות שפותחו בדוח זה.
- (ג) **הפקת פלט.** הפקת דוח תקופתי המסכם את התוצאות, ומקפיץ התראות כפי שהוגדר בשגרת העבודה דלעיל.
- עם הזמן אפשר יהיה ללמוד את המערכת ולשפר אותה בשני מישורים: כוונן עדין (Fine Tuning) של גבולות הבקרה והכללים (Rules) לקבלת התראות מדויקות יותר מחד, ופיתוח השיטות הקיימות ושיטות חדשות מאידך.

### מגבלות המחקר

המגבלה המרכזית במחקר זה הייתה הקושי האינהרנטי בבקרת איכות על ציוני מבחנים המועברים לקבוצות אוכלוסייה קטנות במועדים רבים. במחקר הנוכחי השתמשנו בשיטות ניטור ציונים לנתונים מסוג זה שפותחו בספרות ועל ידינו על מנת להתגבר ולהתמודד עם אופי הנתונים המיוחד.

כמו כן, המידע אודות המשקולות שהוקצו לנוסחים השונים לא נשמר בעבר באופן שוטף. כיום מידע זה נשמר לשגרת בדיקה עתידית. בנוסף, לא תמיד יש לנו נתונים דמוגרפיים מלאים לגבי הנבחנים, למשל נתון ה"מגדר" המשמש לניבוי במשוואת הרגרסיה, מאחר ומדובר בשדה רשות בטופס הבחינה.

### מחקר עתידי – מימ"ד ובחינות נוספות

- נושאים נוספים שנרצה לבחון בהמשך מחקר זה ובמחקרים עתידיים:
- **שיפור בחינת סדר הפרקים.** בחינת האפשרות לשפר את הבדיקה הנוכחית על ידי עריכת מבחני מובהקות על כל פרק בנפרד, ולא לכל הפרקים גם יחד, כפי שנעשה במחקר זה.
  - **קישור לבחינת הפסיכומטרי.** בחינת הקשר בין הציון הפסיכומטרי לציון בבחינת המימ"ד, ושימוש בממצאים לבקרות.
  - **שיפור ניבוי ציון הבחינה הכללי באמצעות הרגרסיה ליניארית.** שיפור טיב הניבוי, למשל, על ידי הוספת חודש הבחינה כמשתנה דמה שיהווה מנבא נוסף; עריכת רגרסיה היררכית במקום מרובה לאמידת התרומה הסגולית של כל אחד מהמנבאים.

- **קיבוץ המכללות לצורך ניבוי.** קיבוץ המכללות השונות שבהן נערך המבחן על פי אופיין באמצעות ציון המימ"ד ו/הפסיכומטרי של הנבחנים, ושימוש בנתון זה כחזאי נוסף במשוואת הרגרסיה. נעשה ניסיון ראשוני לעשות כן באמצעות ניתוח אשכולות, אולם טרם עלה בידינו לקבץ את המוסדות באשכולות מובחנים מספיק וברי משמעות.

**לסיכום,** המחקר בוצע על נתוני עבר של בחינת המימ"ד, מתוך מטרה לבקר באופן מתמשך את הציונים בזמן אמת, ולפתח טכניקות בקרת איכות חדשות לבקרה עתידית. המחקר ממשיך מתוך כוונה לשפר את בדיקות בקרת האיכות שנעשו, ליישמן כשגרת בקרות אוטומטית, ולערוך בקרת איכות בפורמט זה גם על בחינות ליניאריות ואדפטיביות אחרות.

## מקורות

- אללוף, א' ופרונטון, מ' (2012). נתונים על אודות בקרת איכות בבחינת מימ"ד. דוח פנימי שלא פורסם, מאל"ו.
- באומר, מ' גפני, נ' טורוול א' ושביד צ' (2010). *לוח סיכום בחינת מילד לשנים 2008-2009*. דו"ח פנימי שלא פורסם, מאל"ו.
- המרכז הארצי לבחינות ולהערכה (2013). אתר המרכז הארצי לבחינות ולהערכה. אוהזר ב-1.1.2013, מתוך: <https://www.nite.org.il/index.php/he/tests.html>
- Allalouf, A. (1999) *Scoring and Equating at the National Institute for Testing & Evaluation*. Research Report no. 269, Jerusalem: NITE.
- Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, 26(1), 36-46.
- Deming, W. E. (1982). *Out of the crisis*. Cambridge, MA: MIT Press.
- Gutentag, T. (October, 2013). "Quality Control Measures for Scoring Continuously Administered Tests". Paper presented at the 39th Annual IAEA Conference, Tel-Aviv, Israel.
- Kolen, M. J., & Brennan, R. L. (2004). Chapter 8: Practical issues in equating. *Test equating: Linking and scaling* (pp. 268-328). New York: Springer.
- Lee, Y., & von Davier, A. A. (2013). Monitoring Scale Scores over Time via Quality Control Charts, Model-Based Approaches, and Time Series Techniques. *Psychometrica*.
- Montgomery, D. C. (2009). *Introduction to statistical quality control, sixth edition*. Hoboken, NJ: John Wiley & Sons.
- Schafer, W. D., Coverdale, B. J., Luxenberg, H., & Jin, Y. (2011). Quality Control Charts in Large-Scale Assessment Programs. *Practical Assessment, Research & Evaluation*, 16(15). [Available online: <http://pareonline.net/getvn.asp?v=16&n=15>.
- Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand.
- Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. W.E. Deming (Ed.). Washington DC: Department of Agriculture.
- Wild, C. L., and Ramaswamy, R. (Eds.) (2008). *Improving Testing: Applying Process Tools and Techniques to Assure Quality*. New York: Lawrence Erlbaum Associates.

## נספח 1 – שיטות לבקרת איכות

**Harmonic Regression** – רגרסיה שבה המשתנים המנבאים הינם פונקציות טריגונומטריות (ולכן מחזוריות) של זמן. רגרסיה כזו מאפשרת בניית מודל לנתונים עונתיים.

**Cumulative Sum (CUSUM) Control Charts** – מיועדת לאיתור מהיר של שינוי קבוע, אפילו לא מאוד בולט, בערכו הממוצע של מדד מסוים. כל נתון שמתקבל מושווה לנתון קיצוני (נמוך ביותר או גבוה ביותר) שהתקבל עד כה. המדד עבור נקודת זמן מסוימת הוא החרגה המצטברת ביחס לערך מטרה מסוים. כשהחרגה עולה על סף מוגדר, ניתן לזהות שינוי מגמתי בנתונים.

**Time Series Models** – הניתוח מיועד למטרה כפולה: (1) ניתוח חד מימדי של נתונים שמתקבלים לאורך זמן, דבר המאפשר לאתר מגמות ועונתיות, לנכות אותם, וכך לזהות חריגה מערך הצפוי. ישנו עניין מיוחד בקשר בין נתונים בנקודות זמן ספציפית לבין נתונים בנקודות הזמן הסמוכות אליה. (2) ניתוח נתוני פאנל, המאפשר לגלות קשרים בין משתנה הזמן למשתנים אחרים תלויי זמן.

**Change-Point Models** – רגרסיה המיועדת לבחון מהי נקודת הזמן שבה ההתפלגות של הנתונים משתנה. מתאימה מודלים מהתפלגויות שונות לקטעי זמן שונים בנתונים, כאשר השאלה בה מתעניינים היא מהן "נקודות המעבר" בין ההתפלגויות.

**Mixed-Model ANOVA** – שיטה בה ניתן לבחון גורמים קטגוריאליים העשויים להשפיע על הבדלים בין מדידות שנעשו בנקודות זמן שונות, ולאחר אינטראקציות בין הגורמים המשפיעים לבין עצמם, ובין הגורמים המשפיעים לזמן המדידה.

## נספח 2 – התפלגות נוסחי הבחינה בכול חודש באחוזים

| נוסה באחוזים |    |    |    |    |    |    |    |    |      | חודש-שנה |
|--------------|----|----|----|----|----|----|----|----|------|----------|
| F1           | D2 | C2 | B2 | A1 | E1 | D1 | C1 | B1 | N    |          |
| 0            | 0  | 0  | 0  | 0  | 0  | 46 | 0  | 54 | 78   | 2008-05  |
| 0            | 0  | 0  | 0  | 0  | 0  | 52 | 0  | 48 | 531  | 2008-06  |
| 0            | 0  | 0  | 0  | 0  | 0  | 51 | 0  | 49 | 866  | 2008-07  |
| 0            | 0  | 0  | 0  | 0  | 0  | 49 | 0  | 51 | 1303 | 2008-08  |
| 0            | 0  | 0  | 0  | 0  | 0  | 47 | 0  | 53 | 1292 | 2008-09  |
| 0            | 0  | 0  | 0  | 0  | 0  | 57 | 0  | 43 | 188  | 2008-10  |
| 0            | 0  | 0  | 0  | 0  | 0  | 51 | 0  | 49 | 366  | 2008-12  |
| 0            | 0  | 0  | 0  | 0  | 0  | 43 | 0  | 57 | 791  | 2009-01  |
| 0            | 0  | 0  | 0  | 0  | 0  | 40 | 0  | 60 | 688  | 2009-02  |
| 0            | 0  | 0  | 0  | 0  | 0  | 37 | 0  | 63 | 78   | 2009-03  |
| 0            | 0  | 0  | 0  | 0  | 50 | 22 | 0  | 28 | 186  | 2009-05  |
| 0            | 0  | 0  | 0  | 0  | 26 | 66 | 0  | 8  | 449  | 2009-06  |
| 0            | 0  | 0  | 0  | 0  | 42 | 44 | 0  | 15 | 987  | 2009-07  |
| 0            | 0  | 0  | 0  | 0  | 45 | 21 | 14 | 19 | 1358 | 2009-08  |
| 0            | 0  | 0  | 0  | 0  | 64 | 4  | 25 | 8  | 1353 | 2009-09  |
| 0            | 0  | 0  | 0  | 0  | 55 | 24 | 21 | 0  | 227  | 2009-10  |
| 0            | 0  | 0  | 0  | 0  | 69 | 19 | 12 | 0  | 216  | 2009-11  |
| 0            | 0  | 0  | 0  | 0  | 56 | 38 | 6  | 0  | 491  | 2009-12  |
| 0            | 0  | 0  | 0  | 0  | 75 | 14 | 11 | 0  | 367  | 2010-01  |
| 0            | 0  | 0  | 0  | 0  | 77 | 16 | 8  | 0  | 450  | 2010-02  |
| 0            | 0  | 0  | 0  | 0  | 79 | 15 | 6  | 0  | 107  | 2010-03  |
| 0            | 0  | 0  | 0  | 0  | 86 | 8  | 6  | 0  | 36   | 2010-04  |
| 0            | 16 | 16 | 20 | 23 | 9  | 3  | 5  | 6  | 128  | 2010-05  |
| 0            | 16 | 26 | 28 | 22 | 10 | 0  | 0  | 0  | 367  | 2010-06  |
| 0            | 14 | 21 | 26 | 27 | 12 | 0  | 0  | 0  | 618  | 2010-07  |
| 0            | 31 | 13 | 13 | 12 | 31 | 0  | 0  | 0  | 1636 | 2010-08  |
| 0            | 31 | 12 | 10 | 14 | 34 | 0  | 0  | 0  | 846  | 2010-09  |
| 0            | 27 | 14 | 11 | 13 | 35 | 0  | 0  | 0  | 626  | 2010-10  |
| 0            | 24 | 10 | 6  | 12 | 48 | 0  | 0  | 0  | 123  | 2010-11  |
| 0            | 33 | 13 | 13 | 13 | 29 | 0  | 0  | 0  | 318  | 2010-12  |
| 0            | 35 | 13 | 15 | 13 | 24 | 0  | 0  | 0  | 791  | 2011-01  |
| 0            | 35 | 13 | 12 | 12 | 28 | 0  | 0  | 0  | 481  | 2011-02  |
| 0            | 21 | 13 | 16 | 14 | 38 | 0  | 0  | 0  | 96   | 2011-03  |
| 18           | 9  | 9  | 14 | 23 | 27 | 0  | 0  | 0  | 22   | 2011-04  |
| 6            | 13 | 12 | 18 | 38 | 13 | 0  | 0  | 0  | 77   | 2011-05  |
| 13           | 15 | 14 | 13 | 27 | 18 | 0  | 0  | 0  | 520  | 2011-06  |
| 14           | 13 | 14 | 15 | 29 | 14 | 0  | 0  | 0  | 564  | 2011-07  |
| 20           | 19 | 10 | 12 | 22 | 18 | 0  | 0  | 0  | 1244 | 2011-08  |
| 24           | 28 | 13 | 12 | 2  | 21 | 0  | 0  | 0  | 1206 | 2011-09  |
| 25           | 23 | 12 | 13 | 0  | 27 | 0  | 0  | 0  | 436  | 2011-10  |
| 23           | 22 | 16 | 12 | 0  | 27 | 0  | 0  | 0  | 95   | 2011-11  |
| 25           | 28 | 12 | 12 | 0  | 23 | 0  | 0  | 0  | 457  | 2011-12  |
| 25           | 25 | 13 | 13 | 0  | 24 | 0  | 0  | 0  | 834  | 2012-01  |
| 24           | 24 | 13 | 14 | 0  | 25 | 0  | 0  | 0  | 617  | 2012-02  |
| 24           | 16 | 18 | 13 | 0  | 29 | 0  | 0  | 0  | 38   | 2012-03  |

### נספח 3 – אופן חישוב הגבולות

השתמשו ב-X-Bars בלבד

#### Control Limits for $\bar{X}$ - and R-charts

JMP generates control limits for  $\bar{X}$  - and R-charts as follows.

$$\text{LCL for } \bar{X} \text{ chart} = \bar{X}_{\text{w}} - \frac{k\hat{\sigma}}{\sqrt{n_i}}$$

$$\text{UCL for } \bar{X} \text{ chart} = \bar{X}_{\text{w}} + \frac{k\hat{\sigma}}{\sqrt{n_i}}$$

$$\text{LCL for R-chart} = \max(d_2(n_i)\hat{\sigma} - kd_3(n_i)\hat{\sigma}, 0)$$

$$\text{UCL for R-chart} = d_2(n_i)\hat{\sigma} + kd_3(n_i)\hat{\sigma}$$

Center line for R-chart: By default, the center line for the  $i^{\text{th}}$  subgroup, where  $k$  is the number of subgroups, indicates an estimate of the expected value of  $R_i$ , which is computed as  $d_2(n_i)\hat{\sigma}$ , where  $\hat{\sigma}$  is an estimate of  $\sigma$ . If you specify a known value ( $\sigma_0$ ) for  $\sigma$ , the central line indicates the value of  $d_2(n_i)\sigma_0$ . Note that the central line varies with  $n_i$ .

The standard deviation of an  $\bar{X}$  /R chart is estimated by

$$\hat{\sigma} = \frac{\frac{R_1}{d_2(n_1)} + \dots + \frac{R_N}{d_2(n_N)}}{N}$$

where

$\bar{X}_{\text{w}}$  = weighted average of subgroup means

$\sigma$  = process standard deviation

$n_i$  = sample size of  $i^{\text{th}}$  subgroup

$d_2(n)$  is the expected value of the range of  $n$  independent normally distributed variables with unit standard deviation

$d_3(n)$  is the standard error of the range of  $n$  independent observations from a normal population with unit standard deviation

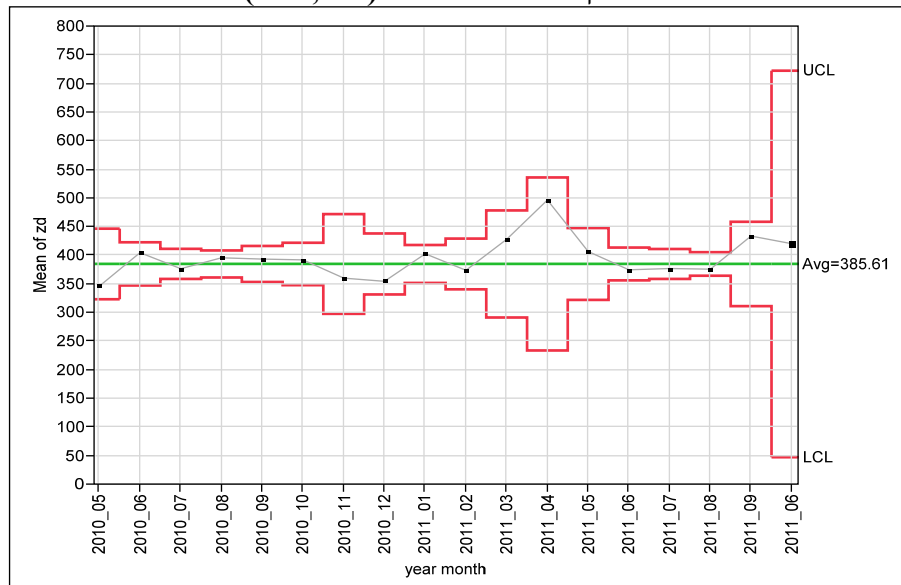
$N$  is the number of subgroups for which  $n_i \geq 2$

## נספח 4 – "מחוץ לשליטה"

| תחום                    | חודש-שנה | N    | ממוצע  | סטיית תקן | UCL    | LCL    | מעל UCL או מתחת LCL? |
|-------------------------|----------|------|--------|-----------|--------|--------|----------------------|
| אנגלית                  | 2009-03  | 78   | 94.03  | 19.99     | 90.02  | 74.26  | above                |
|                         | 2010-03  | 107  | 101.76 | 17.03     | 88.87  | 75.42  | above                |
|                         | 2010-04  | 36   | 95.83  | 17.05     | 93.74  | 70.54  | above                |
|                         | 2010-08  | 1636 | 83.94  | 23.86     | 83.86  | 80.42  | above                |
|                         | 2011-04  | 22   | 103.91 | 12.39     | 96.98  | 67.31  | above                |
|                         | 2008-06  | 531  | 78.78  | 23.29     | 85.16  | 79.12  | below                |
|                         | 2008-09  | 1292 | 80.10  | 22.17     | 84.08  | 80.21  | below                |
|                         | 2010-10  | 626  | 78.46  | 23.51     | 84.92  | 79.36  | below                |
|                         | 2011-08  | 1244 | 79.90  | 23.79     | 84.12  | 80.17  | below                |
|                         | 2011-11  | 95   | 73.88  | 16.23     | 89.28  | 75.00  | below                |
| מתמטיקה                 | 2008-05  | 78   | 90.51  | 22.86     | 89.36  | 74.93  | above                |
|                         | 2009-03  | 78   | 93.32  | 25.61     | 89.36  | 74.93  | above                |
|                         | 2009-05  | 186  | 87.01  | 22.14     | 86.82  | 77.47  | above                |
|                         | 2010-03  | 107  | 107.06 | 16.05     | 88.30  | 75.99  | above                |
|                         | 2010-04  | 36   | 100.50 | 18.69     | 92.76  | 71.53  | above                |
|                         | 2010-09  | 846  | 84.44  | 21.79     | 84.34  | 79.96  | above                |
|                         | 2011-03  | 96   | 94.44  | 23.43     | 88.65  | 75.64  | above                |
|                         | 2012-03  | 38   | 100.71 | 15.43     | 92.48  | 71.81  | above                |
|                         | 2008-08  | 1303 | 79.98  | 20.29     | 83.91  | 80.38  | below                |
|                         | 2009-08  | 1358 | 80.26  | 20.58     | 83.87  | 80.42  | below                |
|                         | 2010-02  | 450  | 78.43  | 21.61     | 85.15  | 79.14  | below                |
|                         | 2011-07  | 564  | 79.19  | 19.84     | 84.83  | 79.46  | below                |
|                         | 2011-08  | 1244 | 79.13  | 20.73     | 83.95  | 80.34  | below                |
|                         | 2011-11  | 95   | 75.43  | 17.89     | 88.68  | 75.61  | below                |
| עברית                   | 2009-09  | 1353 | 82.12  | 20.00     | 81.95  | 78.71  | above                |
|                         | 2009-10  | 227  | 85.84  | 21.07     | 84.28  | 76.37  | above                |
|                         | 2010-03  | 107  | 92.70  | 16.75     | 86.09  | 74.56  | above                |
|                         | 2010-08  | 1636 | 81.93  | 19.61     | 81.80  | 78.85  | above                |
|                         | 2010-10  | 626  | 83.16  | 20.96     | 82.71  | 77.94  | above                |
|                         | 2011-03  | 96   | 92.71  | 20.49     | 86.41  | 74.24  | above                |
|                         | 2011-09  | 1206 | 82.47  | 20.53     | 82.04  | 78.61  | above                |
|                         | 2012-03  | 38   | 91.16  | 17.59     | 90.00  | 70.66  | above                |
|                         | 2008-08  | 1303 | 76.96  | 18.52     | 81.98  | 78.67  | below                |
|                         | 2009-08  | 1358 | 77.64  | 18.60     | 81.94  | 78.71  | below                |
|                         | 2009-11  | 216  | 73.53  | 23.64     | 84.38  | 76.27  | below                |
|                         | 2010-05  | 128  | 70.89  | 22.18     | 85.60  | 75.06  | below                |
|                         | 2010-07  | 618  | 76.63  | 20.32     | 82.72  | 77.93  | below                |
|                         | 2010-11  | 123  | 71.04  | 24.38     | 85.70  | 74.95  | below                |
|                         | 2011-08  | 1244 | 77.71  | 19.45     | 82.02  | 78.64  | below                |
|                         | 2011-11  | 95   | 62.11  | 17.79     | 86.44  | 74.21  | below                |
|                         | 2011-12  | 457  | 77.43  | 21.77     | 83.12  | 77.54  | below                |
| ציון<br>הבחינה<br>הכללי | 2009-03  | 78   | 455.59 | 123.01    | 442.51 | 369.97 | above                |
|                         | 2010-03  | 107  | 519.18 | 77.81     | 437.20 | 375.27 | above                |
|                         | 2010-04  | 36   | 479.39 | 91.81     | 459.62 | 352.85 | above                |
|                         | 2010-08  | 1636 | 414.47 | 104.89    | 414.16 | 398.32 | above                |
|                         | 2011-03  | 96   | 461.03 | 106.75    | 438.93 | 373.55 | above                |
|                         | 2011-04  | 22   | 485.14 | 74.53     | 474.53 | 337.95 | above                |
|                         | 2012-03  | 38   | 485.63 | 82.84     | 458.20 | 354.28 | above                |
|                         | 2008-08  | 1303 | 391.21 | 100.96    | 415.11 | 397.36 | below                |
|                         | 2009-08  | 1358 | 395.88 | 102.10    | 414.93 | 397.55 | below                |
|                         | 2009-11  | 216  | 382.47 | 118.86    | 428.03 | 384.44 | below                |
|                         | 2011-08  | 1244 | 390.06 | 105.26    | 415.32 | 397.16 | below                |
|                         | 2011-11  | 95   | 336.57 | 86.08     | 439.10 | 373.37 | below                |

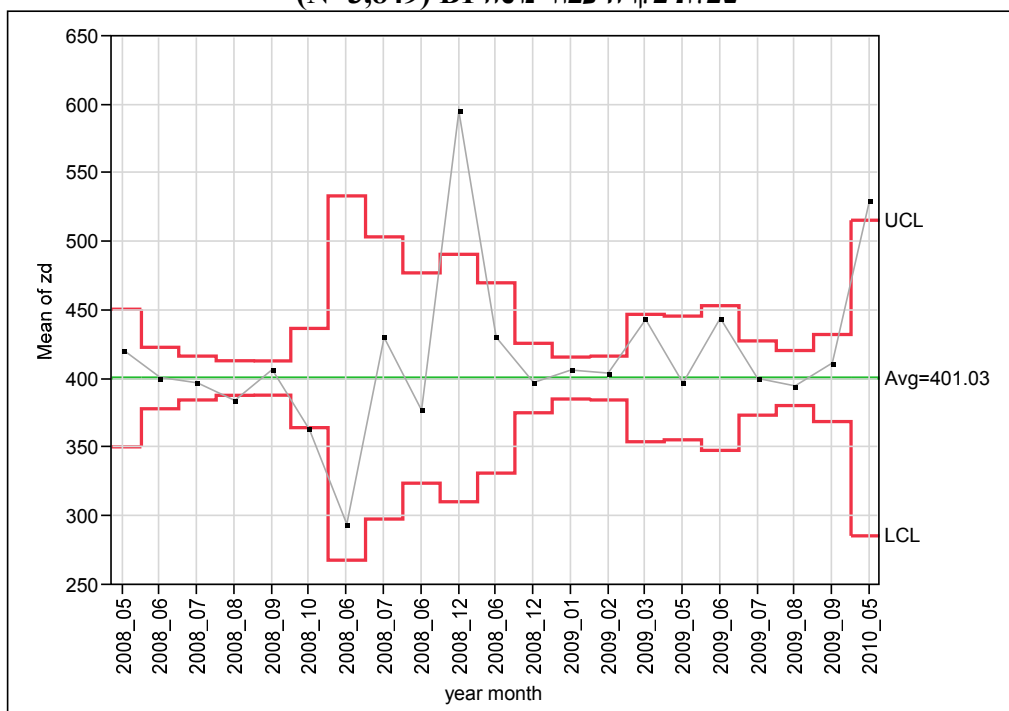
## נספח 5 – טבלאות בקרה לכל נוסח בנפרד

טבלת בקרה עבור נוסח A1 (N=1,542)



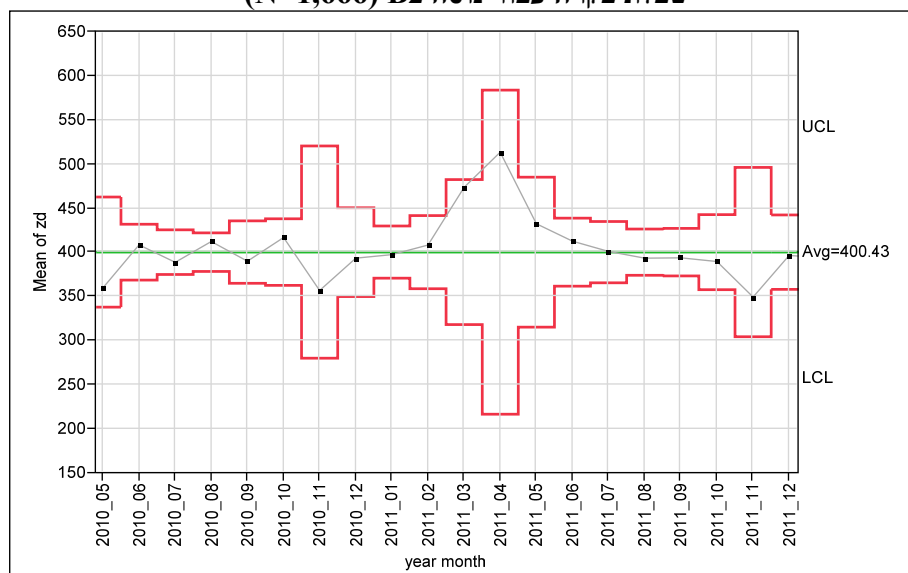
| חודש-שנה | N   | ממוצע ציון הבחינה הכללי |
|----------|-----|-------------------------|
| 2010_05  | 30  | 346.8000                |
| 2010_06  | 80  | 405.2625                |
| 2010_07  | 165 | 377.3879                |
| 2010_08  | 204 | 396.3676                |
| 2010_09  | 115 | 394.1043                |
| 2010_10  | 83  | 392.7711                |
| 2010_11  | 15  | 360.8667                |
| 2010_12  | 40  | 355.2000                |
| 2011_01  | 104 | 403.3173                |
| 2011_02  | 58  | 374.5172                |
| 2011_03  | 13  | 428.6923                |
| 2011_04  | 5   | 497.4000                |
| 2011_05  | 29  | 407.2414                |
| 2011_06  | 141 | 375.6099                |
| 2011_07  | 166 | 377.5301                |
| 2011_08  | 273 | 376.6374                |
| 2011_09  | 21  | 434.3810                |

טבלת בקרה עבור נוסח B1 (N=3,849)



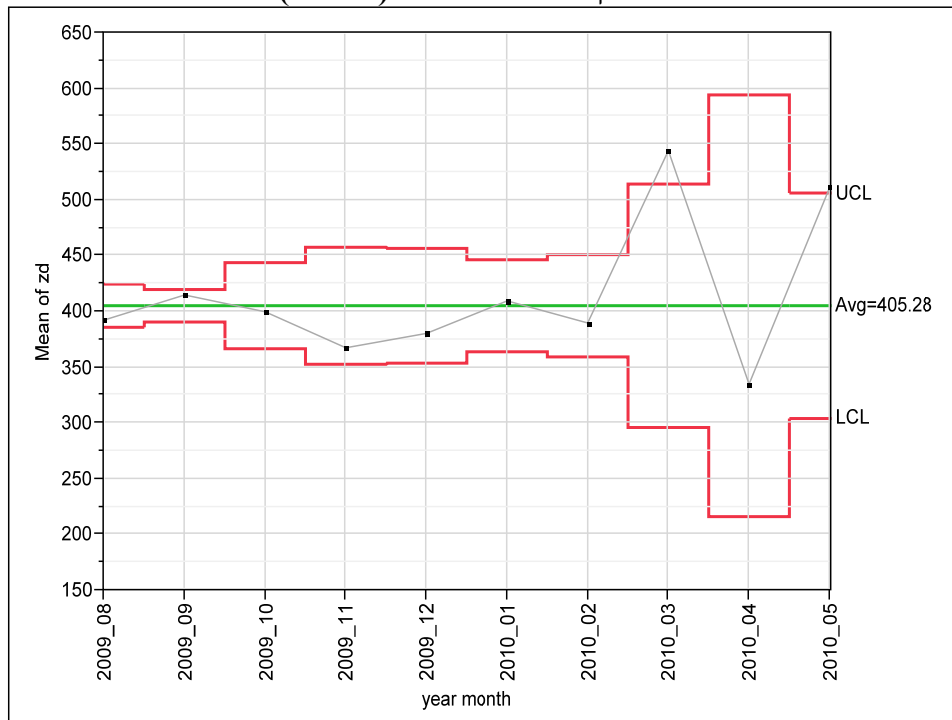
| ממוצע ציון הבחינה הכללי | N   | חודש-שנה |
|-------------------------|-----|----------|
| 421.8095                | 42  | 2008_05  |
| 399.8210                | 257 | 2008_06  |
| 398.3153                | 425 | 2008_07  |
| 384.6541                | 662 | 2008_08  |
| 407.162                 | 679 | 2008_09  |
| 363.3457                | 81  | 2008_10  |
| 412.1742                | 178 | 2008_12  |
| 406.8524                | 454 | 2009_01  |
| 404.6223                | 413 | 2009_02  |
| 443.7755                | 49  | 2009_03  |
| 397.4231                | 52  | 2009_05  |
| 444.6053                | 38  | 2009_06  |
| 400.7431                | 144 | 2009_07  |
| 395.3206                | 262 | 2009_08  |
| 411.9524                | 105 | 2009_09  |
| 530.6250                | 8   | 2010_05  |

**טבלת בקרה עבור נוסח B2 (N=1,666)**



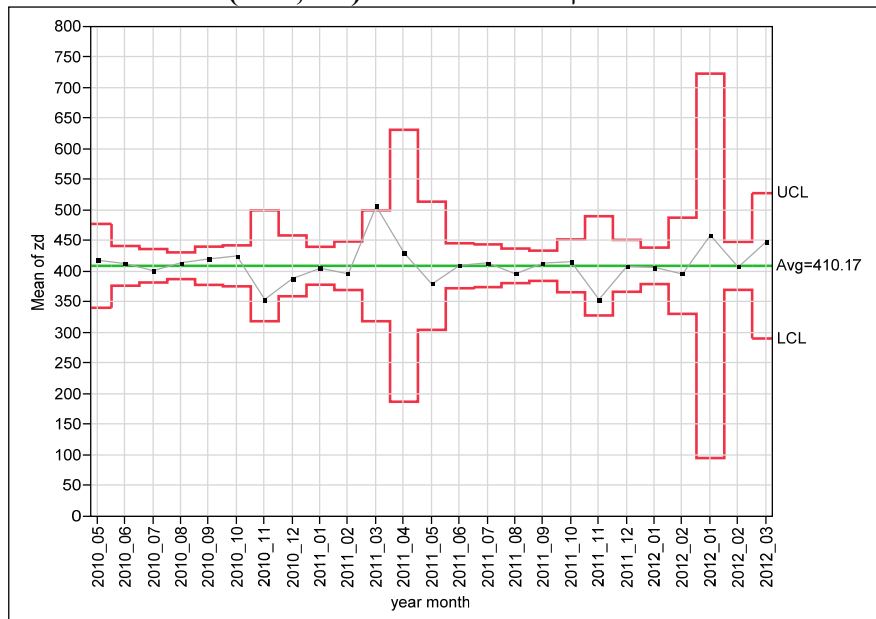
| ממוצע ציון הבחינה הכללי | N   | חודש-שנה |
|-------------------------|-----|----------|
| 359.8846                | 26  | 2010_05  |
| 408.6337                | 101 | 2010_06  |
| 388.9114                | 158 | 2010_07  |
| 412.4245                | 212 | 2010_08  |
| 390.2346                | 81  | 2010_09  |
| 416.9437                | 71  | 2010_10  |
| 357.0000                | 7   | 2010_11  |
| 393.6000                | 40  | 2010_12  |
| 397.5086                | 116 | 2011_01  |
| 408.4407                | 59  | 2011_02  |
| 473.8667                | 15  | 2011_03  |
| 513.6667                | 3   | 2011_04  |
| 432.5714                | 14  | 2011_05  |
| 412.8824                | 68  | 2011_06  |
| 401.2976                | 84  | 2011_07  |
| 393.6803                | 147 | 2011_08  |
| 393.9500                | 140 | 2011_09  |
| 389.8750                | 56  | 2011_10  |
| 349.6364                | 11  | 2011_11  |
| 396.6667                | 57  | 2011_12  |
| 394.0183                | 109 | 2012_01  |
| 405.0465                | 86  | 2012_02  |
| 471.2000                | 5   | 2012_03  |

טבלת בקרה עבור נוסח C1 (N=720)



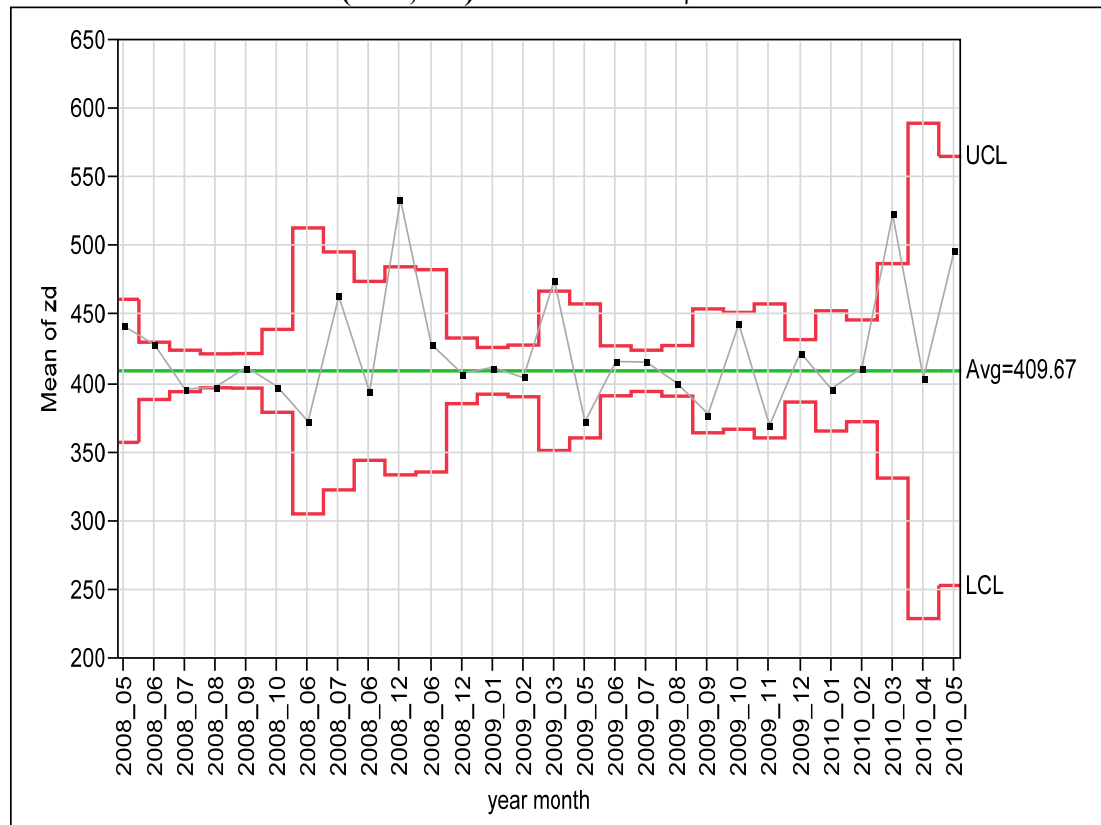
| חודש-שנה | N   | ממוצע ציון הבחינה הכללי |
|----------|-----|-------------------------|
| 2009_08  | 191 | 392.4660                |
| 2009_09  | 337 | 414.9496                |
| 2009_10  | 48  | 399.9167                |
| 2009_11  | 26  | 367.5769                |
| 2009_12  | 27  | 380.4815                |
| 2010_01  | 42  | 409.6429                |
| 2010_02  | 34  | 389.8235                |
| 2010_03  | 6   | 544.3333                |
| 2010_04  | 2   | 335.0000                |
| 2010_05  | 7   | 511.1429                |

**טבלת בקרה עבור נוסח C2 (N=1,645)**



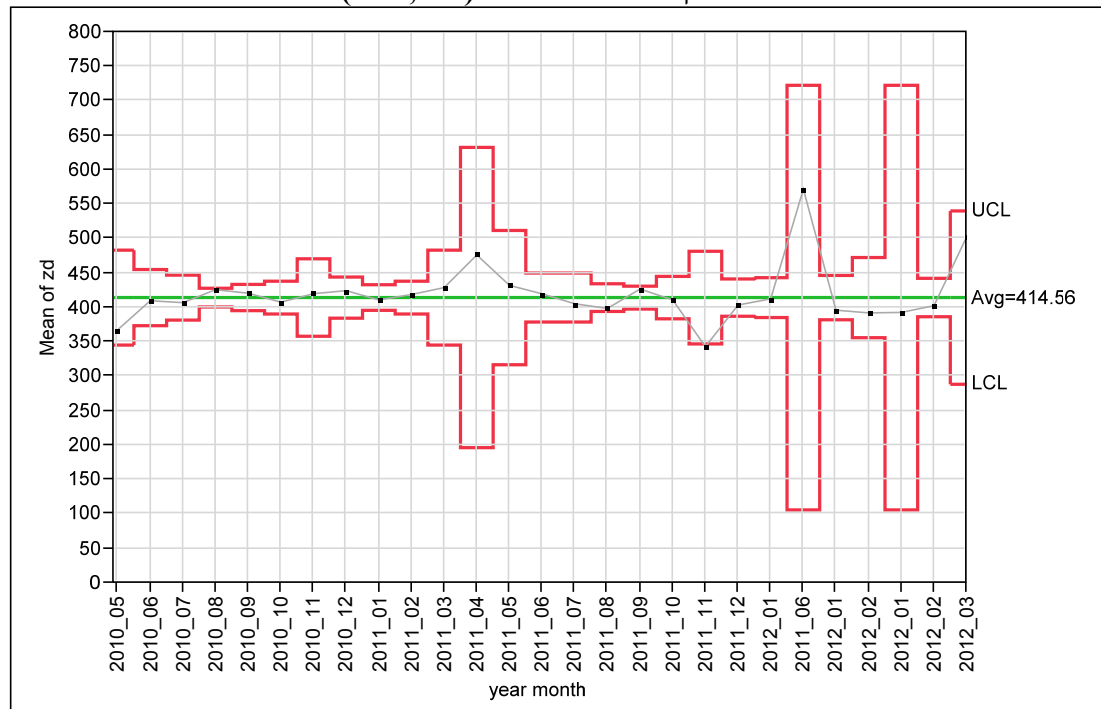
| ממוצע ציון הבחינה הכללי | N   | חודש-שנה |
|-------------------------|-----|----------|
| 419.3810                | 21  | 2010_05  |
| 413.8617                | 94  | 2010_06  |
| 402.3864                | 132 | 2010_07  |
| 415.2271                | 207 | 2010_08  |
| 421.4455                | 101 | 2010_09  |
| 426.2386                | 88  | 2010_10  |
| 355.2500                | 12  | 2010_11  |
| 389.4000                | 40  | 2010_12  |
| 406.3107                | 103 | 2011_01  |
| 397.0317                | 63  | 2011_02  |
| 507.7500                | 12  | 2011_03  |
| 431.5000                | 2   | 2011_04  |
| 380.2222                | 9   | 2011_05  |
| 410.7808                | 73  | 2011_06  |
| 415.0988                | 81  | 2011_07  |
| 397.0242                | 124 | 2011_08  |
| 414.5776                | 161 | 2011_09  |
| 416.9811                | 53  | 2011_10  |
| 355.2000                | 15  | 2011_11  |
| 409.0182                | 55  | 2011_12  |
| 407.8929                | 112 | 2012_01  |
| 405.6750                | 80  | 2012_02  |
| 449.4286                | 7   | 2012_03  |

טבלת בקרה עבור נוסח D1 (N=4,476)



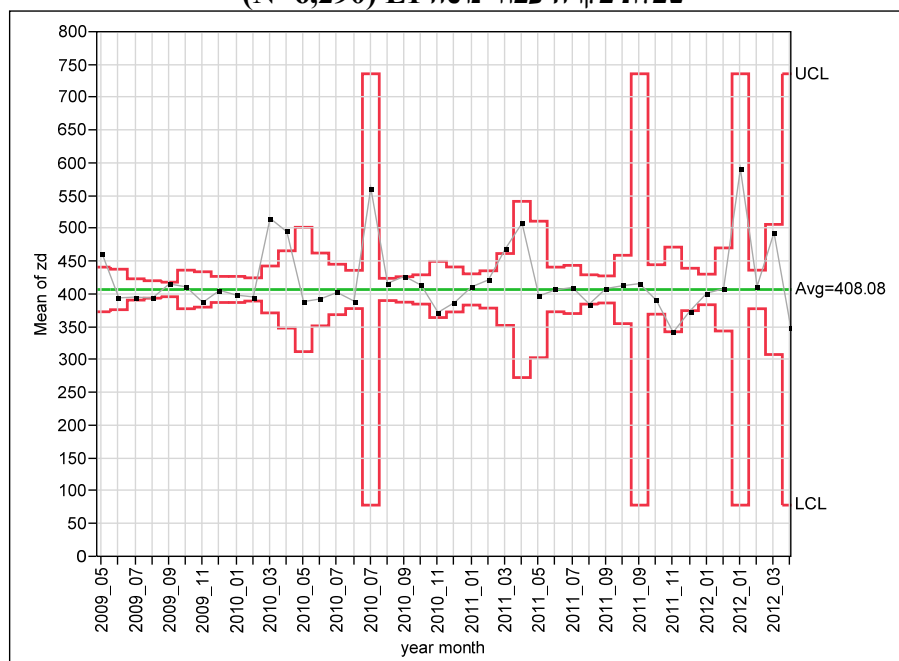
| ממוצע ציון הבחינה הכללי | N   | חודש-שנה |
|-------------------------|-----|----------|
| 442.3611                | 36  | 2008_05  |
| 424.2117                | 274 | 2008_06  |
| 398.2812                | 441 | 2008_07  |
| 397.9766                | 641 | 2008_08  |
| 412.4812                | 613 | 2008_09  |
| 398.2150                | 107 | 2008_10  |
| 419.2447                | 188 | 2008_12  |
| 412.3056                | 337 | 2009_01  |
| 404.8000                | 275 | 2009_02  |
| 475.5517                | 29  | 2009_03  |
| 372.6098                | 41  | 2009_05  |
| 416.4696                | 296 | 2009_06  |
| 416.1316                | 433 | 2009_07  |
| 400.6701                | 288 | 2009_08  |
| 378.2708                | 48  | 2009_09  |
| 443.7963                | 54  | 2009_10  |
| 370.0244                | 41  | 2009_11  |
| 423.1376                | 189 | 2009_12  |
| 396.5882                | 51  | 2010_01  |
| 411.9437                | 71  | 2010_02  |
| 523.3125                | 16  | 2010_03  |
| 404.3333                | 3   | 2010_04  |
| 496.7500                | 4   | 2010_05  |

טבלת בקרה עבור נוסח D2 (N=3,062)



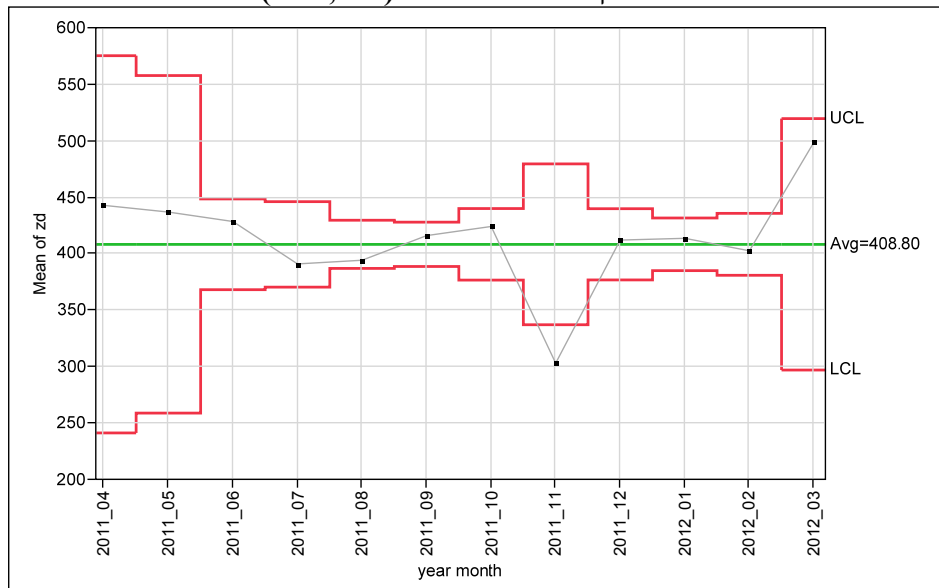
| חודש-שנה | N   | ממוצע ציון הבחינה הכללי |
|----------|-----|-------------------------|
| 2010_05  | 20  | 367.7500                |
| 2010_06  | 57  | 409.9298                |
| 2010_07  | 89  | 407.1798                |
| 2010_08  | 514 | 426.0603                |
| 2010_09  | 262 | 420.9008                |
| 2010_10  | 167 | 407.8623                |
| 2010_11  | 30  | 420.1333                |
| 2010_12  | 106 | 424.6226                |
| 2011_01  | 276 | 411.1377                |
| 2011_02  | 167 | 419.0180                |
| 2011_03  | 20  | 429.0000                |
| 2011_04  | 2   | 476.5000                |
| 2011_05  | 10  | 432.2000                |
| 2011_06  | 76  | 421.9868                |
| 2011_07  | 75  | 405.6133                |
| 2011_08  | 236 | 398.9280                |
| 2011_09  | 342 | 427.0409                |
| 2011_10  | 100 | 411.1000                |
| 2011_11  | 21  | 343.3810                |
| 2011_12  | 129 | 403.6047                |
| 2012_01  | 207 | 405.3913                |
| 2012_02  | 150 | 400.8267                |
| 2012_03  | 6   | 502.8333                |

# טבלת בקרה עבור נוסח E1 (N=6,290)



| ממוצע ציון הבחינה הכללי | N   | חודש-שנה |
|-------------------------|-----|----------|
| 463.0323                | 93  | 2009_05  |
| 394.9391                | 115 | 2009_06  |
| 396.5732                | 410 | 2009_07  |
| 394.9449                | 617 | 2009_08  |
| 416.3882                | 863 | 2009_09  |
| 412.1680                | 125 | 2009_10  |
| 388.4966                | 149 | 2009_11  |
| 406.6182                | 275 | 2009_12  |
| 399.4453                | 274 | 2010_01  |
| 396.8087                | 345 | 2010_02  |
| 516.6235                | 85  | 2010_03  |
| 495.9677                | 31  | 2010_04  |
| 389.7500                | 12  | 2010_05  |
| 393.6000                | 35  | 2010_06  |
| 405.7432                | 74  | 2010_07  |
| 410.4870                | 499 | 2010_08  |
| 428.2509                | 287 | 2010_09  |
| 414.5714                | 217 | 2010_10  |
| 373.0508                | 59  | 2010_11  |
| 388.3913                | 92  | 2010_12  |
| 412.0156                | 192 | 2011_01  |
| 422.5746                | 134 | 2011_02  |
| 469.5833                | 36  | 2011_03  |
| 509.3333                | 6   | 2011_04  |
| 398.2000                | 10  | 2011_05  |
| 407.5269                | 93  | 2011_06  |
| 411.0750                | 80  | 2011_07  |
| 385.7306                | 219 | 2011_08  |
| 409.3504                | 254 | 2011_09  |
| 399.1810                | 116 | 2011_10  |
| 342.6923                | 26  | 2011_11  |
| 375.3077                | 104 | 2011_12  |
| 403.2200                | 200 | 2012_01  |
| 412.4079                | 152 | 2012_02  |
| 494.7273                | 11  | 2012_03  |

טבלת בקרה עבור נוסח F1 (N=1,298)



| חודש-שנה | N   | ממוצע ציון הבחינה הכללי |
|----------|-----|-------------------------|
| 2011_04  | 4   | 443.2500                |
| 2011_05  | 5   | 437.4000                |
| 2011_06  | 69  | 429.0145                |
| 2011_07  | 78  | 391.3846                |
| 2011_08  | 245 | 394.6367                |
| 2011_09  | 288 | 416.5625                |
| 2011_10  | 111 | 424.6126                |
| 2011_11  | 22  | 303.5909                |
| 2011_12  | 112 | 412.6161                |
| 2012_01  | 206 | 414.1068                |
| 2012_02  | 149 | 403.0000                |
| 2012_03  | 9   | 499.2222                |

## נספח 6 – מקדמי מתאם רגרסיה ליניארית

| P-value | t ערך  | מקדם מתוקנן (β) | טעות תקן | מקדמי מתאם (B) | תתי מנבאים            |
|---------|--------|-----------------|----------|----------------|-----------------------|
| <.0001* | 231.65 | 0.00            | 1.68     | 389.61         | Intercept             |
| <.0001* | -32.88 | -0.23           | 0.74     | -24.32         | gender[f]             |
| <.0001* | 14.75  | 0.11            | 1.49     | 22.02          | time_multiplier[1.0]  |
| 0.58    | -0.55  | 0.00            | 1.70     | -0.94          | time_multiplier[1.25] |
| 0.0125* | 2.50   | 0.03            | 9.72     | 24.29          | campus[CA]            |
| <.0001* | 5.04   | 0.04            | 2.94     | 14.83          | campus[RA]            |
| <.0001* | -5.01  | -0.04           | 3.43     | -17.16         | campus[SA]            |
| <.0001* | -6.33  | -0.05           | 2.90     | -18.38         | campus[VA]            |
| <.0001* | 6.12   | 0.06            | 6.69     | 40.93          | campus[GB]            |
| 0.0015* | -3.17  | -0.03           | 6.57     | -20.87         | campus[IB]            |
| 0.0003* | -3.65  | -0.04           | 6.98     | -25.48         | campus[MB]            |
| 0.0086* | -2.63  | -0.02           | 4.43     | -11.64         | campus[YD]            |
| 0.5341  | -0.62  | -0.01           | 6.12     | -3.81          | campus[CE]            |
| 0.0063* | 2.73   | 0.02            | 3.96     | 10.80          | campus[IE]            |
| <.0001* | 6.52   | 0.06            | 6.63     | 43.21          | campus[JE]            |
| 0.1161  | -1.57  | -0.01           | 4.76     | -7.47          | campus[LE]            |
| <.0001* | -4.63  | -0.04           | 3.93     | -18.18         | campus[MG]            |
| 0.2297  | -1.20  | -0.01           | 9.41     | -11.30         | campus[RG]            |
| 0.0255* | 2.23   | 0.02            | 3.47     | 7.76           | campus[CH]            |
| 0.0428* | 2.03   | 0.02            | 5.05     | 10.23          | campus[DH]            |
| 0.0005* | -3.46  | -0.03           | 5.85     | -20.24         | campus[JH]            |
| <.0001* | 8.49   | 0.06            | 2.99     | 25.41          | campus[UH]            |
| <.0001* | 20.20  | 0.17            | 5.01     | 101.27         | campus[OJ]            |
| <.0001* | -7.77  | -0.06           | 3.43     | -26.61         | campus[CK]            |
| 0.004*  | -2.88  | -0.02           | 5.11     | -14.71         | campus[IK]            |
| <.0001* | 7.05   | 0.07            | 6.81     | 48.02          | campus[OK]            |
| <.0001* | 18.53  | 0.15            | 3.75     | 69.44          | campus[CL]            |
| 0.0151* | -2.43  | -0.05           | 20.18    | -49.03         | campus[OL]            |
| 0.7991  | -0.25  | 0.00            | 2.52     | -0.64          | campus[VL]            |
| 0.3868  | -0.87  | -0.01           | 3.92     | -3.39          | campus[CN]            |
| 0.0041* | 2.87   | 0.02            | 4.65     | 13.35          | campus[BO]            |
| 0.005*  | -2.81  | -0.06           | 21.17    | -59.50         | campus[HO]            |
| <.0001* | -11.40 | -0.14           | 9.78     | -111.49        | campus[MO]            |
| <.0001* | -11.44 | -0.10           | 5.83     | -66.65         | campus[TO]            |
| 0.1753  | 1.36   | 0.01            | 4.76     | 6.45           | campus[NR]            |
| 0.0051* | 2.80   | 0.03            | 5.96     | 16.71          | campus[PR]            |
| <.0001* | -4.72  | -0.04           | 3.30     | -15.57         | campus[PS]            |
| 0.078   | 1.76   | 0.01            | 4.92     | 8.68           | campus[SS]            |
| <.0001* | 4.44   | 0.04            | 5.40     | 23.97          | campus[HT]            |
| <.0001* | 13.11  | 0.13            | 7.04     | 92.29          | campus[IT]            |
| <.0001* | -4.01  | -0.03           | 3.93     | -15.74         | campus[VI]            |

\*  $p < .05$