



מרכז ארצי לבחינות ולהערכה (ע"ר)  
NATIONAL INSTITUTE FOR TESTING & EVALUATION  
المركز القطري للامتحانات والتقييم  
מיסודן של האוניברסיטאות בישראל

## 21 שנות בחינת אמיר"ם - בקרת איכות ומחקר



דצמבר 2015

טוני גוטנטיג | אבי אללוף | מיכל באומר | נדב פודולר | מרינה פרונטון

דוח מאלו 420

ISBN:978-965-502-192-9

## **21 שנות בחינת אמיר"ס – בקרת איכות ומחקר**

טוני גוטנטג, אבי אללוף, מיכל באומר, נדב פודולר ומרינה פרונטון

דוח מאל"ו

דצמבר 2015

## תוכן עניינים

|         |                                              |
|---------|----------------------------------------------|
| 3.....  | תקציר                                        |
| 4.....  | ABSRTACT                                     |
| 5.....  | מבוא                                         |
| 6.....  | מבחן אמיר"ם                                  |
| 10..... | חלק א' – סיכום 20 שנות העברה של בחינת אמיר"ם |
| 14..... | חלק ב' – בקרת איכות                          |
| 26..... | סיכום                                        |
| 28..... | מקורות                                       |

## טבלאות

|         |                                                                                                                                                       |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| 6.....  | טבלה 1 - טבלת השוואה בין בחינות ידע באנגלית במאל"ו                                                                                                    |
| 7.....  | טבלה 2 - אחוזי התכנסות לפי נוסח 2012-2013                                                                                                             |
| 10..... | טבלה 3 - שכיחות, ממוצע וסטיית תקן של הציון ו- $\Theta$ לפי שנה                                                                                        |
| 10..... | טבלה 4 - מגדר - ממוצע וסטיית תקן ציון ו- $\Theta$ , 2003-2013                                                                                         |
| 14..... | טבלה 5 - שכיחות, ממוצע וסטיית תקן של הציון ו- $\Theta$ לפי נוסח 1992-2013                                                                             |
| 15..... | טבלה 6 - שכיחות, ממוצע וסטיית תקן, של הציון וה- $\Theta$ לפי גרסאות נוסחים                                                                            |
| 19..... | טבלה 7 - פוסט-הוק מסוג HSD על ציון הבחינה הכללי 2012-2013                                                                                             |
| 22..... | טבלה 8 - ממוצע (ס.ת.) ומספר נבחנים בקבוצת נבחנים לפי מספר בחינה 1992-2013                                                                             |
| 23..... | טבלה 9 - ממוצע, סטיית תקן ומספר נבחנים לפי מספר היבחנות 1992-2013                                                                                     |
| 23..... | טבלה 10 - ממוצע וסטיית תקן של הפרשי הציונים והפרש מתוקנן בין ההיבחנות ומספר הנבחנים 1992-2013                                                         |
| 23..... | טבלה 11 - מספר הנבחנים, ממוצע, סטיית תקן וחציון של מרווחי הזמן (בימים) בין היבחנות 1992-2013                                                          |
| 24..... | טבלה 12 - מספר הנבחנים שקיבלו כל נוסח לפי נוסח ומספר הפעמים שהנוסח התקבל 1992-2013                                                                    |
| 25..... | טבלה 13 - שיעור הנבחנים שקיבלו נוסח זהה או שונה בהעברה חוזרת, ממוצע הפרשי הציונים (ס.ת.), הפרש מתוקנן, בחינת מובהקות ושכיחות לפי מספר העברה 1992-2013 |

## תרשימים

|         |                                                                     |
|---------|---------------------------------------------------------------------|
| 8.....  | תרשים 1 - הקשר בין $\Theta$ (תטא) לציון לפי שנה                     |
| 11..... | תרשים 2 - מגדר – ציון ממוצע לפי שנה 2003-2013                       |
| 11..... | תרשים 3 - גיל - שכיחות וציון ממוצע מעבר לשנים 2003-2013             |
| 12..... | תרשים 4 - שכיחות נבחנים על-פי מוסד אקדמי מעבר לשנים 1992-2013       |
| 12..... | תרשים 5 - שכיחות הנבחנים באוניברסיטה הפתוחה לעומת יתר המוסדות       |
| 13..... | תרשים 6 - שכיחות הנבחנים באמיר"ם ובאמיר"ר לפי שנה                   |
| 13..... | תרשים 7 - ממוצע ציוני הנבחנים באמיר"ם ובאמיר"ר לפי שנה              |
| 16..... | תרשים 8 - ציון ממוצע על-פני ציר הזמן לפי נוסחים 1992-2013           |
| 17..... | תרשים 9 - ציון ממוצע על-פני ציר הזמן לפי גרסת נוסח, 1992-2003       |
| 18..... | תרשים 10 - ציון ממוצע על-פני ציר הזמן לפי גרסת נוסח, 2004-2013      |
| 19..... | תרשים 11 - צפיפות: התפלגות הציון לפי נוסח 2012-2013                 |
| 20..... | תרשים 12 - תרשימי בקרה על הציונים 2012-2013                         |
| 21..... | תרשים 13 - תרשימי בקרה על שאריות הציונים 2012-2013                  |
| 22..... | תרשים 14 - ממוצעי הציונים בקרב הנבחנים החוזרים                      |
| 24..... | תרשים 15 - התפלגות הפרשי זמני ההעברות בימים 1992-2013               |
| 27..... | תרשים 16 - צילום מסך מערכת ממוכנת לבקרה מתמשכת על בחינת המימ"ד 2014 |

עניינו של המחקר הנוכחי הוא בקרת איכות בחינת אמיר"ם, שהיא בחינה באופנות מתמשכת. זוהי בחינה ממוחשבת אדפטיבית המשמשת למיון לרמות שליטה בשפה האנגלית עבור מוסדות לימוד אקדמיים, הנבנית ומועברת על ידי המרכז הארצי לבחינות ולהערכה (מאל"ו). מ-1992 עד 2013 התקיימו באמיר"ם 199,674 היבחנויות ב-11 נוסחים. בחינת אמיר"ם מועברת יותר משני עשורים, ומחקר זה מאפשר גם התבוננות "היסטורית" ארוכת טווח על התפלגות הציונים.

**סיכום 21 שנות העברה של בחינת אמיר"ם.** לצד העלייה העקבית במספר הנבחנים במשך השנים, נרשמה במקביל ירידה בציון הממוצע בבחינה. יש שימוש גובר בבחינה במוסדות לימוד חדשים לצד הוותיקים. הגיל הממוצע של הנבחנים הוא כ-27, מתוכם, 43% הם גברים, וממוצע ההישגים שלהם גבוה מזה של הנשים. החל משנת 2004, מספר הנבחנים בבחינת אמיר"ם גבוה ממספר הנבחנים בבחינת אמיר"ר – בחינה ליניארית המועברת לאותה מטרה – והציון הממוצע באמיר"ם נמוך מהציון הממוצע באמיר"ר.

**בקרת איכות.** בקרת האיכות שנערכה במחקר זה על בחינת אמיר"ם התבססה על שיטות סטטיסטיות שנחקרו ופותחו לאחרונה במאל"ו כדי לפקח על בחינות המועברות באופנות מתמשכת. הציונים בנוסחי הבחינה השונים יציבים למדי, וזאת במיוחד בעשור האחרון. יש הבדל מובהק אך קטן בין ממוצע הציונים של הנוסחים השונים שהועברו ב-2012-2013. תרשימי בקרת האיכות המציגים השתנות מדדים סטטיסטיים מרכזיים לאורך זמן הראו כי תהליך הציון הינו תקין. כרבע מהבחניות שנערכו היו בחינות חוזרות, והפרש הימים החציוני בין העברות עוקבות עומד על כחצי שנה. ההפרש בין ממוצעי הציונים של נבחנים שנבחנו באותו נוסח בחינה (אותו מאגר פריטים) בשתי העברות עוקבות היה גבוה באופן מובהק מזה של נבחנים שנבחנו בנוסחים מקבילים, אך האפקט היה זעיר.

המחקר הנוכחי מרחיב את היריעה לקראת גיבוש תפיסה כוללת של בקרת איכות על מבחנים המועברים תכופות לקבוצות נבחנים קטנות, ונותן מענה ראשוני לבקרה על הציונים המופקים במבחנים אדפטיביים בנוסף למענה הקיים למבחנים לינאריים. אחד מתוצרי המחקר הינו הרחבת פעילותה של המערכת הממוכנת לבקרה שגרתית של הציונים, המופעלת כיום על בחינת מימ"ד, גם לבחינת אמיר"ם. מחקר עתידי יתמקד בהכללה של שגרת בקרת האיכות שפותחה במחקר הנוכחי גם על בחינות אחרות.

## ABSRTACT

This research focuses on Quality Control (QC) for AMIRAM, a Continuous Mode Test (CMT) used by academic institutions for English-language placement of students. AMIRAM is a computerized adaptive test that uses a 3-parameter IRT. The database at NITE was used in the current investigation. Between the years 1992-2013, AMIRAM was administered 199,674 times, using 11 test form versions. Based on over two decades of AMIRAM testing, this study thus offers a long-view historical perspective on the data.

**A 21-year summary of the AMIRAM test.** Alongside the steady increase in the number of examinees over the years, there has been a decrease in the mean test score. The examinees' mean age was 27 and, on average, males (43%) got higher scores than female examinees. The use of the test in new academic institutions has been on the rise, without any decrease in its use in other institutions. The number of AMIRAM examinees is greater than that of AMIR examinees, and the mean test score of the AMIRAM is lower compared to the AMIR.

**Quality Control.** The quality control procedures used for the present study rely on recently developed statistical methods employed in the context of CMT. The scores in the different AMIRAM test forms have been quite stable, especially in the last decade. There was a statistically significant, albeit small, difference in the test forms' mean score in 2012-2013. A quality control chart presenting a statistical index over time showed that the scoring procedure is relatively "in control". A quarter of the tests were repeat tests, and the median difference in the time between consecutive administrations is about half a year. The difference in mean test score – though very small – is significantly higher for examinees who received the same test form (item bank) in two consecutive administrations, compared with examinees who received a parallel form.

The present study continues to extend the scope of quality control for CMT, and is a first attempt to address the issue of QC for adaptive testing in addition to linear tests. One of the outcomes of this research has been the adaptation of a computerized quality control system from the MEIMAD test to AMIRAM. Future research will focus on validating the QC procedures developed in this study for use in other exams.

בקרת איכות היא הליך חיוני הן עבור הגופים העורכים מבחנים, הן עבור הגופים המשתמשים בציונים וכמובן עבור הנבחנים. האחריות המוטלת על ארגון המפתח מבחנים, מעביר אותם, מחשב ציונים, ומדווח אותם לנבחנים ולמוסדות, הינה לדאוג לכך שהציון יהיה מדויק, מהימן ותקף. על-מנת להבטיח זאת, עליו לעקוב אחרי- ולבקר את- תהליכי פיתוח המבחן, העברת המבחן, חישוב הציונים ודיווחם באופן שגורתי. הביקורת תאפשר לארגון לאתר טעויות שנעשו בעבר, לנטר טעויות בהווה, ולמנוע טעויות בעתיד.

מבחינים בשני מצבים של בקרת איכות: בקרת איכות של בחינות המועברות לקבוצות גדולות במספר קטן של מועדים קבועים תוך שימוש במספר נוסחים קטן (לרוב מדובר ב- Large Scale Assessment; למשל, הפסיכומטרי, SAT, ACT); ובקרת איכות של בחינות המועברות לקבוצות אוכלוסייה קטנות במספר מועדים רב תוך שימוש בנוסחי בחינה רבים ("העברה באופנות מתמשכת", CMT – Continuous Mode Tests; למשל, אמיר"ם, מימ"ד, TOEFL, IELTS). שיטות הבקרה במצב הראשון ידועות ומוכרות (Allalouf, 2007; Kolen & Brennan, 2004), אך הן אינן מתאימות לצורך בקרה במצב השני וזאת בשל גודלי המדגמים הקטנים ובשל המספר הרב של נוסחים המועברים במקביל.

לאחרונה פותחו בעולם שיטות סטטיסטיות לבקרת איכות של בחינות באופנות העברה מתמשכת (Lee & von Davier, 2013; Schafer et al., 2011). מאל"י פיתח לאחרונה הליך לביצוע בקרת איכות של בחינת המימ"ד (אללוף ועמיתיו, 2014; באומר ועמיתיה, 2014; Allalouf & Gutentag, 2015) ועל בחינת אמיר"ם. שממצאיה יתוארו בדוח זה (גוטנטג ועמיתיה, 2015; Allalouf & Gutentag, under revision; Allalouf et al., 2014; Gutentag et al., 2013).

עניינו של המחקר הנוכחי הוא בקרת איכות של בחינת אמיר"ם המועברת באופנות העברה מתמשכת על בסיס השיטות שפותחו לאחרונה במאל"י ושיטות אחרות. מטרת המחקר לבנות שגרת בדיקות של בקרת איכות לבחינת אמיר"ם, באמצעות כלי ממוחשב (באומר, בהכנה). מאחר שבחינת אמיר"ם מועברת כבר יותר משני עשורים (1992-2013), המחקר מאפשר גם התבוננות "היסטורית" ארוכת טווח על הציונים ואיכותם.

## מבחן אמיר"ם

בבסיס הנתונים של מאל"י נשמרים הציונים של כל הנבחנים שנבחנו בבחינות שקיימות במאל"י החל משנת 1983. החל מ-1992 נפתחה ההרשמה לאמיר"ם. בבחינה זו ההרשמה נערכת דרך מוסד הלימודים. בין אוגוסט 1992 ל-דצמבר 2013 נבחנו באמיר"ם 199,674 נבחנים (חלקם יותר מפעם אחת).

מאל"י מקיים באופן שגרתי שלוש בחינות מרכזיות הבוחנות ידע באנגלית (כשפה שנייה): פרקי אנגלית בבחינה הפסיכומטרית, בחינת אמיר"ר ובחינת אמיר"ם; ראו טבלה 1 להשוואת המטרה, התכנים ואופנות ההעברה בין שלוש הבחינות.

**טבלה 1 - טבלת השוואה בין בחינות ידע באנגלית במאל"י**

| הבחינה הפסיכומטרית | אמיר"ר                                                      | אמיר"ם                               |                                |
|--------------------|-------------------------------------------------------------|--------------------------------------|--------------------------------|
| <b>מטרה</b>        | כלי לחיזוי הצלחה אקדמית; מיון לחוגי לימוד; מיון רמת אנגלית. | מיון רמת אנגלית בעיקר באוניברסיטאות. | מיון רמת אנגלית בעיקר במכללות. |
| <b>תכנים</b>       | מילולי, כמותי, אנגלית                                       | אנגלית                               | אנגלית                         |
| <b>אופנות</b>      | נייר ועיפרון                                                | נייר ועפרון                          | ממוחשב אדפטיבי                 |

המחקר הנוכחי מתמקד בבחינת אמיר"ם. בחינת אמיר"ם היא גרסה ממוחשבת אדפטיבית של בחינת אמיר"ר. בדומה לתחום האנגלית בבחינה הפסיכומטרית, הבחינה בודקת שליטה פאסיבית בשפה האנגלית ברמה של קריאת טקסטים והבנתם. בבחינה שאלות משלושה סוגים: השלמת משפטים, ניסוח מחדש והבנת הנקרא. השאלות הן שאלות ברירה עם ארבע תשובות אפשריות. משך הבחינה הממוצע כשעה. בחינת אמיר"ם נערכת אחת לחודשיים-שלושה, ואין הגבלת זמן בין בחינות חוזרות.

בחינת אמיר"ם מיועדת לסטודנטים או למועמדים למוסדות אקדמיים בישראל, אשר המוסד בו הם לומדים או נרשמו אליו מבקש לאמוד את ידיעותיהם באנגלית, אם לשם קבלתם ללימודים ואם בכדי למיינס לרמות של קורסי החובה באנגלית או בכדי לפטור אותם מהם. השימוש בציון נעשה בהתאם לשיקוליו של כל מוסד.

לאורך תקופת המחקר, היה שימוש ב-11 נוסחי בחינה בבחינת אמיר"ם, המתחלקים ל-27 גרסאות נוסחים<sup>1</sup>. כל "נוסח" הוא למעשה מאגר, המכיל כ-200 פריטים, מתוכם נבחרים עבור הנבחן 21-28 פריטים. הבחינה אדפטיבית ומבוססת על מודל IRT תלת פרמטרי. התאמת קושי השאלות ליכולת הנבחן מאפשרת לאמוד את רמת היכולת שלו בעזרת מספר שאלות קטן יחסית.

<sup>1</sup> גרסאות של אותו נוסח הינן נוסח העל בשינויים קלים. למשל, D1 ו-D2 הן גרסאות של נוסח D.



## חישוב ציון הבחינה

הפרמטר  $\theta$  המחושב על-סמך מודל IRT לוגיסטי תלת פרמטרי (3PL) מייצג את רמת היכולת של הנבחן, במקרה זה בתחום האנגלית. - טבלאות ההמרה מפיקות עבור כל ערך  $\theta$  (בטווח של 3- עד 3) את ציון התקן המתאים לו (50 עד 150, בהתאמה). טבלת ההמרה נקבעה על סמך מחקרים אמפיריים והחלטות לגבי ערכים ספציפיים.

## התכנסות

P-var הינה שונות טעות המדידה של היכולת, הנאמדת לאחר כל תגובה (דהיינו, תשובת נבחן) לפריט בחינה. טעות המדידה היא ביחס הפוך לאינפורמציה: בעוד שהאינפורמציה הולכת וגדלה ככל שהנבחן משיב על יותר פריטים, טעות המדידה הולכת וקטנה. השאיפה היא למזער את טעות המדידה ככל האפשר. כאשר הטעות יורדת מתחת לסף שהוגדר מראש נאמר כי חלה "התכנסות". לעיתים, המבחן יסתיים ללא התכנסות בנקודת החתך.

בטבלה 2 מוצגים אחוזי התכנסות שהתקבלו לכל נוסח בשנים 2012-2013. סף אחוז ההתכנסות מוגדר באמיר"ם כ-  $P-var \leq 0.08$ . ערך זה משקף מהימנות של 0.92 לנבחן (בדומה למהימנות בחינת אמיר"ר). ניתן לראות מהטבלה כי ממוצע אחוזי ההתכנסות בשנים 2012-2013 יציבים למדי ועומדים על כ-80%.

טבלה 2 - אחוזי התכנסות לפי נוסח 2012-2013

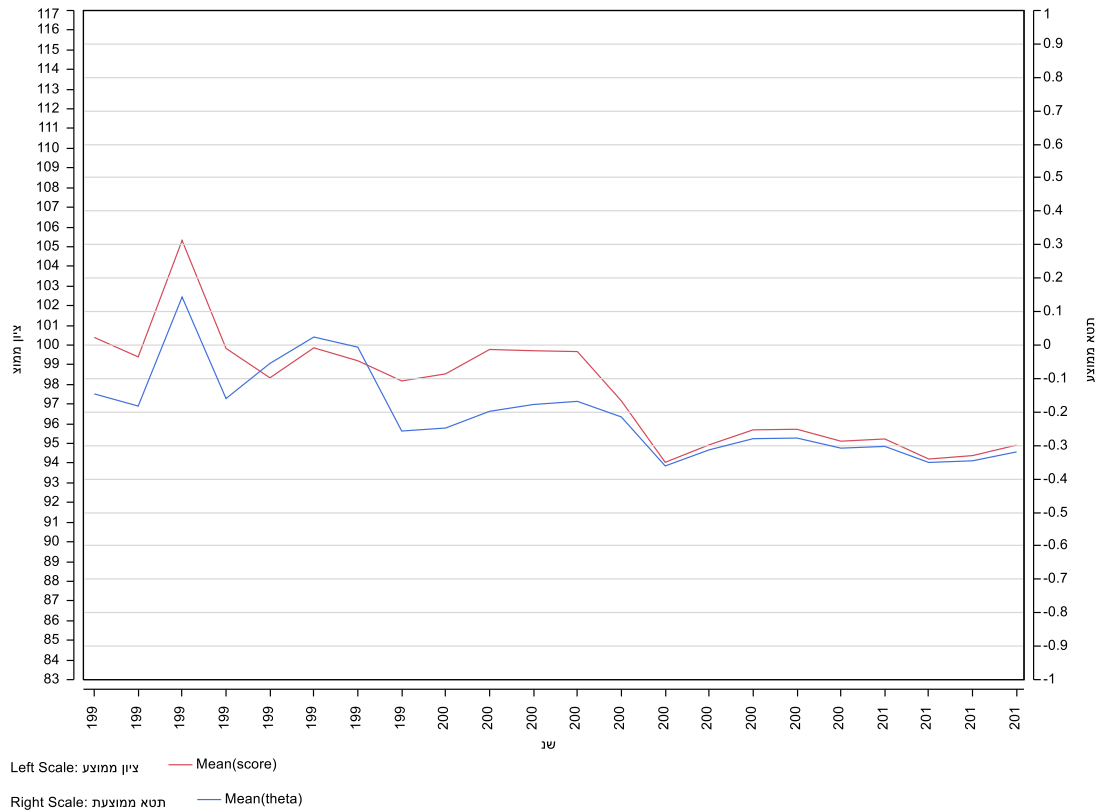
| נוסח | אחוז הנבחנים שציונם התכנס |
|------|---------------------------|
| C    | 80                        |
| D    | 80                        |
| E    | 77                        |
| F    | 80                        |
| G    | 80                        |
| H    | 80                        |
| I    | 81                        |
| J    | 79                        |

מספר הפריטים המינימאלי המוצג לנבחן באמיר"ם הוא 21 והמספר המרבי הוא 28. מהשאלה ה-21, בודקים לאחר כל תגובה לפריט אם יש התכנסות של הציון. במידה ויש התכנסות, המבחן מסתיים, ובמידה שאין ממשיכים עד לפריט ה-28.

## טבלאות המרה

תרשים 1 מציג את הקשר בין ציון  $\theta$  (תטא) לציון הבחינה לפי שנה. ככל שהשנים מתקדמות,  $\theta$  והציון הופכים להיות מתואמים יותר ויותר. מתאם פירסון לבחינת הקשר בין  $\theta$  לציון לאורך השנים הוא כמעט מושלם ( $r = .99$ ).

## תרשים 1 - הקשר בין $\theta$ ((תטא) לציון לפי שנה



דוח זה כולל סיכום 20 שנות העברה של בחינת אמיר"ם ומדדי בקרת איכות על הציונים.

### סיכום 20 שנות העברה של בחינת אמיר"ם

הופקה סטטיסטיקה תיאורית של התפלגות הציונים של ציון הבחינה וציון  $\theta$  לפי שנה (1992-2013). ההתפלגות הציונים הופקה גם בפילוח לפי המשתנים הדמוגרפיים מגדר וגיל, לפי מוסד אקדמי, וכן עבור האוניברסיטה הפתוחה בהשוואה לכל יתר המוסדות. בנוסף, התבצעה השוואה של בחינת אמיר"ם לבחינת אמיר"ר בהתייחס למספר הנבחנים וציוניהם לפי שנה.

### בקרת איכות

בקרת האיכות נבחנה משלושה היבטים: נוסח, תרשימי בקרה ונבחנים חוזרים.

- **נוסח:** הופקו ממוצעי הנוסחים על-פני ציר הזמן, וכן נבחנו הבדלים בין נוסחים.
- **תרשימי בקרה:** הופקו תרשימי בקרה מסוג שוהארט על ממוצע הציונים ועל השאריות מניתוח רגרסיה לאחר ניבוי הציון באמצעות מוסד אקדמי.
- **נבחנים חוזרים:** נבחן השיפור בציון בתוך אותה קבוצות נבחנים וההבדלים בשיפור בין קבוצות נבחנים, הפרשי זמנים בין בחינות חוזרות. כמו כן, נבדקה השאלה האם הנוסחים שנבחנים חוזרים מקבלים שונים בין בחינה לבחינה, ואם לא האם יש לדבר השפעה על הציון.

## חלק א' – סיכום 20 שנות העברה של בחינת אמיר"ם

טבלה 3 מציגה את ההתפלגות של הציון ו- $\theta$  של כלל הנבחנים לפי שנה. ניתן לראות מין הטבלה, כי מספר הנבחנים גדל באופן מונוטוני משנה לשנה. מאידך, הממוצע, הן של הציון והן של  $\theta$ , יורד משנה לשנה, בעוד שבסטיית התקן יש מגמת עליה.

טבלה 3 - שכיחות, ממוצע וסטיית תקן של הציון ו- $\theta$  לפי שנה

| שנה  | N     | ציון   |       | $\theta$ |      |
|------|-------|--------|-------|----------|------|
|      |       | ממוצע  | ס.ת.  | ממוצע    | ס.ת. |
| 1992 | 557   | 100.39 | 20.40 | -0.15    | 1.07 |
| 1993 | 631   | 99.40  | 19.63 | -0.18    | 1.03 |
| 1994 | 1994  | 105.32 | 17.56 | 0.14     | 0.88 |
| 1995 | 2263  | 99.83  | 18.67 | -0.16    | 0.99 |
| 1996 | 2284  | 98.34  | 21.69 | -0.05    | 1.00 |
| 1997 | 3945  | 99.86  | 21.16 | 0.02     | 0.97 |
| 1998 | 4457  | 99.20  | 20.96 | -0.01    | 0.96 |
| 1999 | 5974  | 98.18  | 19.73 | -0.26    | 1.01 |
| 2000 | 5841  | 98.54  | 20.35 | -0.25    | 1.04 |
| 2001 | 6073  | 99.78  | 21.15 | -0.20    | 1.08 |
| 2002 | 7629  | 99.71  | 22.11 | -0.18    | 1.10 |
| 2003 | 9402  | 99.67  | 22.66 | -0.17    | 1.12 |
| 2004 | 9474  | 97.17  | 24.43 | -0.21    | 1.14 |
| 2005 | 10633 | 94.04  | 24.92 | -0.36    | 1.18 |
| 2006 | 12766 | 94.93  | 24.72 | -0.31    | 1.16 |
| 2007 | 14307 | 95.69  | 25.04 | -0.28    | 1.18 |
| 2008 | 15362 | 95.72  | 24.77 | -0.28    | 1.17 |
| 2009 | 14356 | 95.12  | 25.39 | -0.31    | 1.20 |
| 2010 | 15811 | 95.23  | 25.59 | -0.30    | 1.21 |
| 2011 | 15530 | 94.21  | 25.75 | -0.35    | 1.22 |
| 2012 | 18627 | 94.39  | 26.24 | -0.35    | 1.25 |
| 2013 | 21758 | 94.92  | 26.19 | -0.32    | 1.25 |

### מגדר

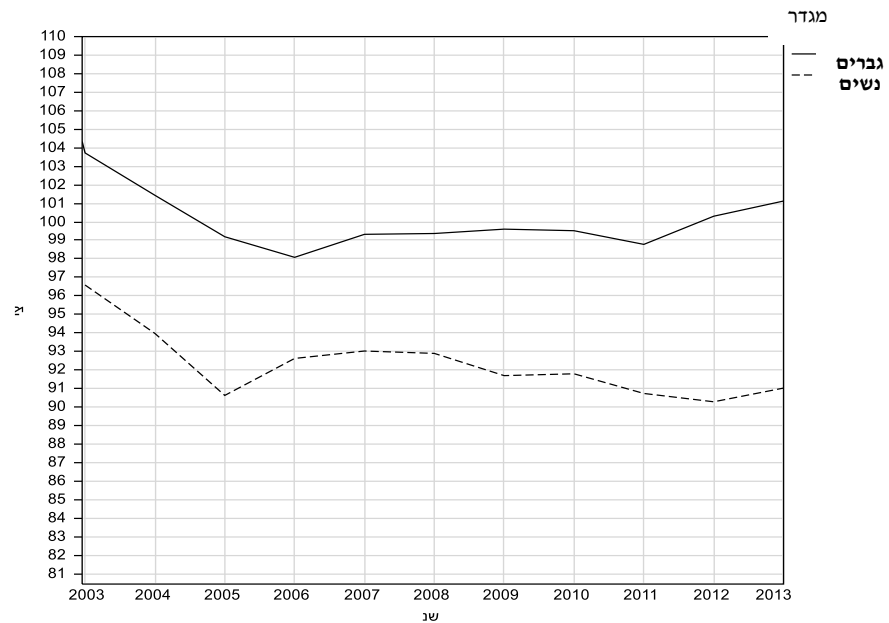
עד שנת 2003 הנבחנים לא נתבקשו לציין את מגדרם, ולאחר שנה זו דיווח על מגדר הינו וולונטרי. מעבר לתקופה, לגבי 43,504 (21.8% מהנבחנים) חסר נתון המגדר. טבלה 4 מציגה ממוצע וסטיית תקן של הציון ו- $\theta$  וכן את סטיות התקן לפי מגדר. שיעור הנשים שנבחנו בבחינה מקרב כלל הנבחנים הוא כ-57%. בממוצע, ציוניהן של הנשים נמוכים במובהק מאלה של הגברים ( $t[156168] = 62.01, p < .001, d = 0.32$ ). בשל ההבדל בשכיחות הנבחנים לפי מגדר, ניתן להניח כי לפחות חלק מההבדל בציון נובע מסלקציה עצמית.

טבלה 4 - מגדר - ממוצע וסטיית תקן ציון ו- $\theta$ , 2003-2013

| מגדר  | N     | ציון  |       | $\theta$ |      |
|-------|-------|-------|-------|----------|------|
|       |       | ממוצע | ס.ת.  | ממוצע    | ס.ת. |
| גברים | 66779 | 99.92 | 25.92 | -0.09    | 1.22 |
| נשים  | 89391 | 91.99 | 24.29 | -0.46    | 1.15 |

בתרשים 2 ניתן לראות ציון ממוצע לפי שנה ולפי מגדר.

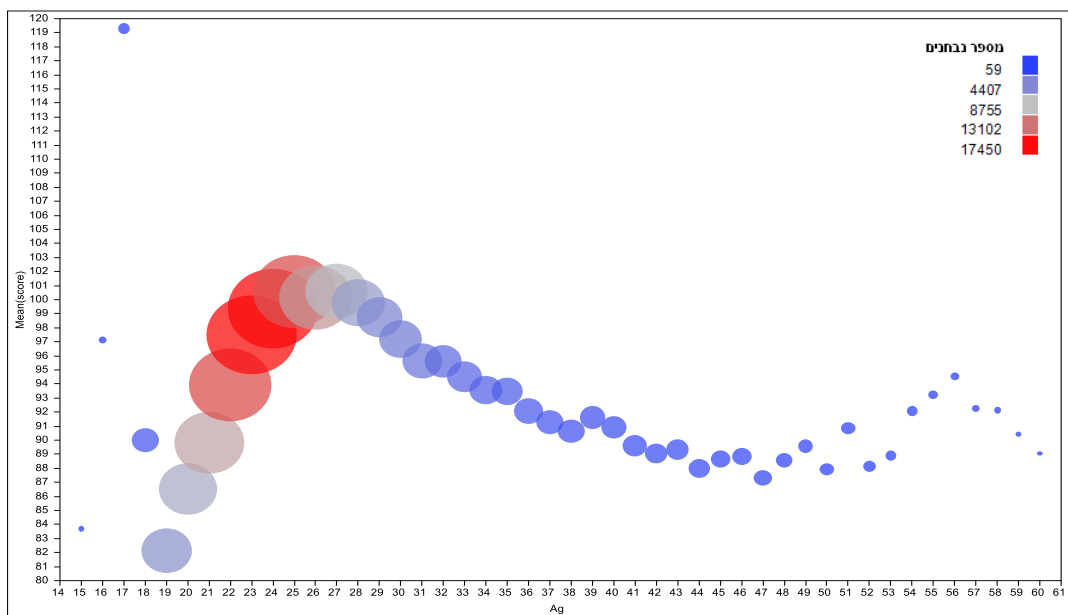
**תרשים 2 - מגדר - ציון ממוצע לפי שנה 2003-2013**



## גיל

עד שנת 2003 הנבחנים לא נתבקשו לציין את גילם, ולאחר שנה זו דיווח על גיל הינו וולונטרי. מעבר לתקופה, לגבי 43,497 (21.8% מהנבחנים) חסר נתון הגיל. בתרשים 3 ניתן לראות את התפלגות הגילאים והציונים מעבר לשנים. ככל שהבועה גדולה יותר וצבעה הופך חם יותר, מספר הנבחנים שהיא משקפת גדול יותר. ישנה התפלגות גילאים בעלת זנב ימני ארוך (ממוצע 26.60, ס.ת. 7.65), כאשר רוב הנבחנים (טווח ביו-רבעוני) הם בני 22-29. הציון הממוצע הגבוה ביותר הוא אצל בני 25.

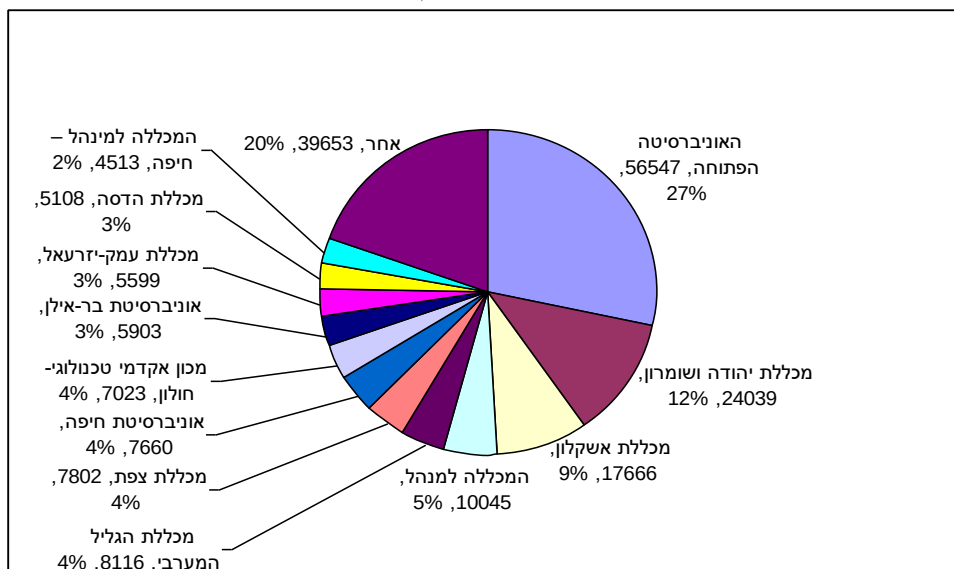
**תרשים 3 - גיל - שכיחות וציון ממוצע מעבר לשנים 2003-2013**



## מוסד אקדמי

תרשים 4 מציג את שכיחות הנבחנים על-פי מוסד אקדמי מעבר לשנים. מוסדות בהם מספר הנבחנים קטן משני אחוז קובצו יחד לקטגורית "אחר". ניתן לראות כי האוניברסיטאות נוטות להשתמש באמיר"ם פחות בהשוואה למכללות. כמו כן, ניתן לראות כי האוניברסיטה הפתוחה, שהיא המוסד הוותיק ביותר שנבחן באמיר"ם, היא בעלת מספר הנבחנים הרב ביותר.

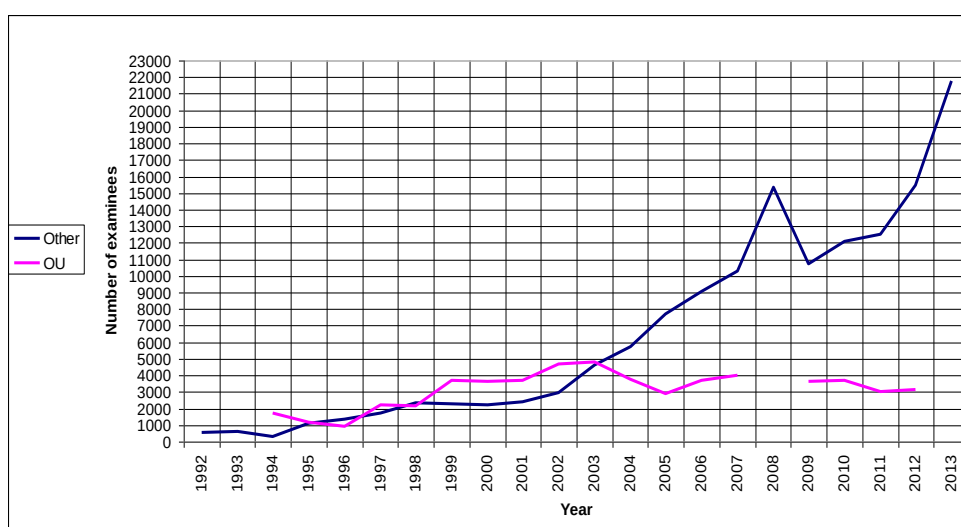
תרשים 4 - שכיחות נבחנים על-פי מוסד אקדמי מעבר לשנים 1992-2013



## האוניברסיטה הפתוחה לעומת יתר המוסדות

תרשים 5 מציג את מספר הנבחנים באוניברסיטה הפתוחה לעומת יתר המוסדות. בעשר השנים הראשונות לקיום הבחינה, מספר הנבחנים מטעם האוניברסיטה הפתוחה היה דומה למספר הנבחנים מטעם כל יתר המוסדות גם יחד. בעוד שמספר הנבחנים באוניברסיטה הפתוחה נותר יחסית קבוע לאורך התקופה, מספר הנבחנים ביתר המוסדות עלה משמעותית עם השנים. לכן, כיום מספר נבחני האוניברסיטה הפתוחה קטן יחסית למספר הנבחנים ביתר המוסדות. בשנים 2008 ו-2013 אין נתונים לגבי השימוש בבחינת אמיר"ם באוניברסיטה הפתוחה.

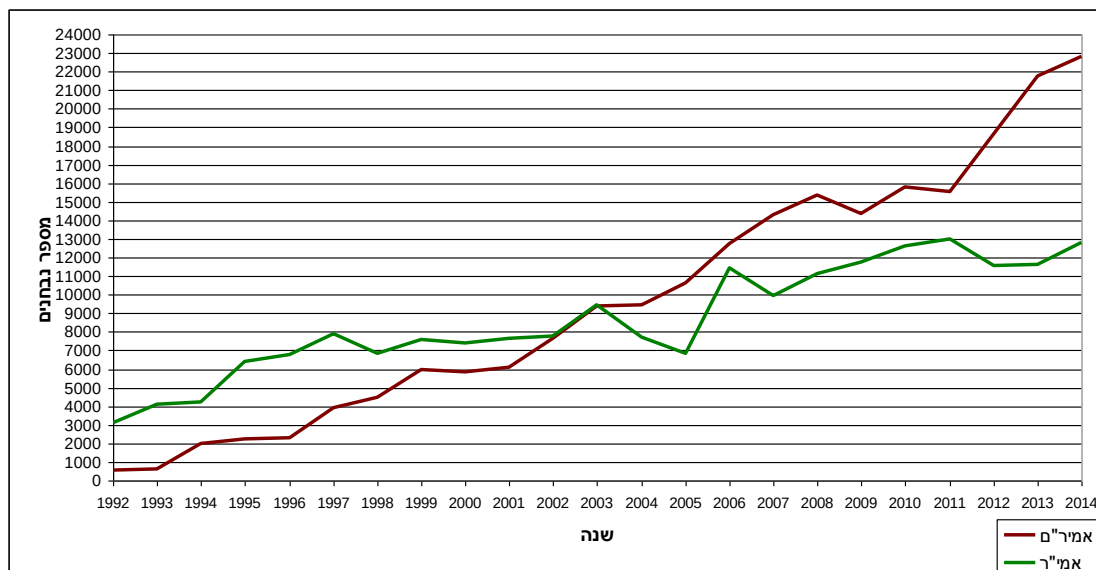
תרשים 5 - שכיחות הנבחנים באוניברסיטה הפתוחה לעומת יתר המוסדות



## בחינות אמיר"ם ואמיר"ר

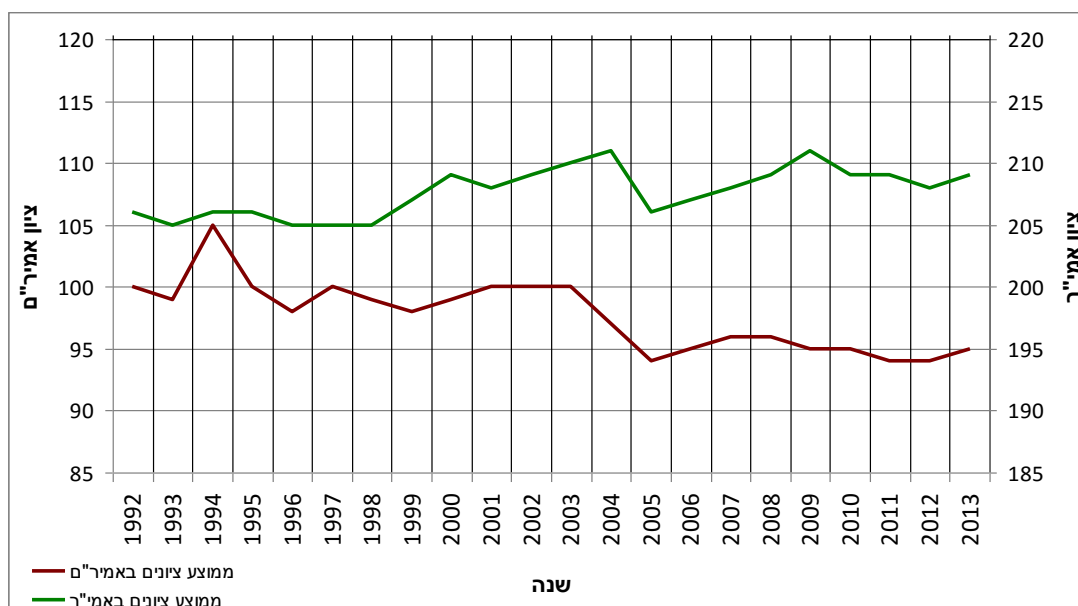
תרשים 6 מציג את שכיחויות הנבחרים באמיר"ם ובאמיר"ר לפי שנה. ניתן לראות כי מספר הנבחרים באמיר"ם עולה בעקביות, ואף עולה על מספר נבחרים באמיר"ר החל משנת 2004.

תרשים 6 - שכיחות הנבחרים באמיר"ם ובאמיר"ר לפי שנה



תרשים 7 מציג את ממוצע ציוני הנבחרים באמיר"ם ובאמיר"ר לפי שנה. כדי להשוות בין ציוני אמיר"ם לאמיר"ר נעשה שימוש בשני צירי y נפרדים: ציוני אמיר"ר ממוקמים על ציר y הימני אשר סולם הציונים שלו גבוה במאה נקודות מסולם ציוני אמיר"ם הממוקם על ציר y השמאלי. ממוצע הציונים בבחינת אמיר"ר גבוה באופן עקבי מממוצע הציונים בבחינת אמיר"ם, כאשר הפער בציונים בין שתי הבחינות הלך וגדל עם השנים, ועמד בשנת 2013 על כ-14 נקודות.

תרשים 7 - ממוצע ציוני הנבחרים באמיר"ם ובאמיר"ר לפי שנה



## חלק ב' – בקרת איכות

בקרת איכות בהקשר של מבחנים מוגדרת כהליך שיטתי פורמאלי המתוכנן להבטיח שתקני איכות גבוהה נשמרים בכל השלבים של הצינון, ניתוח הבחינה ודיווח הצינונים, ובכך ממזערים טעויות ומגדילים את מהימנות המדידה (ITC, 2014). בקרת האיכות תיבחן להלן משלושה היבטים: נוסח, תרשימי בקרה ונבחנים חוזרים.

### נוסח

בבקרת האיכות המתבססת על הנוסחים נבחנה התפלגות הנוסחים והצינונים על-פני ציר הזמן, וכן נבחנו הבדלים בצינונים הממוצעים בין נוסחי הבחינה.

בטבלה 5, הנוסחים מסודרים מהישן ביותר לחדש ביותר. בטבלה מוצגים הממוצע וסטיית התקן של הציון ו- $\theta$ . הנתונים מראים ירידה בציון הממוצע בנוסחים הראשונים, ואילו מנוסח C ואילך ניראה כי ממוצע ציוני הנוסחים וסטיית התקן שלהם הופכים להיות דומים.

טבלה 5 - שכיחות, ממוצע וסטיית תקן של הציון ו- $\theta$  לפי נוסח 1992-2013

| נוסח | N     | ציון הבחינה |       | ציון $\theta$ |      |
|------|-------|-------------|-------|---------------|------|
|      |       | ממוצע       | ס.ת.  | ממוצע         | ס.ת. |
| A    | 16072 | 100.13      | 20.44 | -0.02         | 0.97 |
| BA   | 12996 | 98.25       | 20.63 | -0.27         | 1.06 |
| BB   | 11809 | 99.85       | 21.00 | -0.18         | 1.07 |
| C    | 18049 | 95.80       | 24.40 | -0.30         | 1.14 |
| D    | 20274 | 95.47       | 24.57 | -0.31         | 1.15 |
| E    | 27397 | 95.59       | 24.98 | -0.29         | 1.18 |
| F    | 22165 | 94.73       | 25.70 | -0.32         | 1.23 |
| G    | 17570 | 94.47       | 25.54 | -0.34         | 1.21 |
| H    | 19668 | 95.02       | 25.36 | -0.31         | 1.21 |
| I    | 20149 | 96.02       | 25.63 | -0.27         | 1.22 |
| J    | 13529 | 96.02       | 26.62 | -0.27         | 1.27 |

טבלה 6 מציגה את הנתונים ברזולוציה גבוהה יותר, ברמת גרסאות הנוסחים להבדיל מרמת הנוסחים המאגדים אותם.



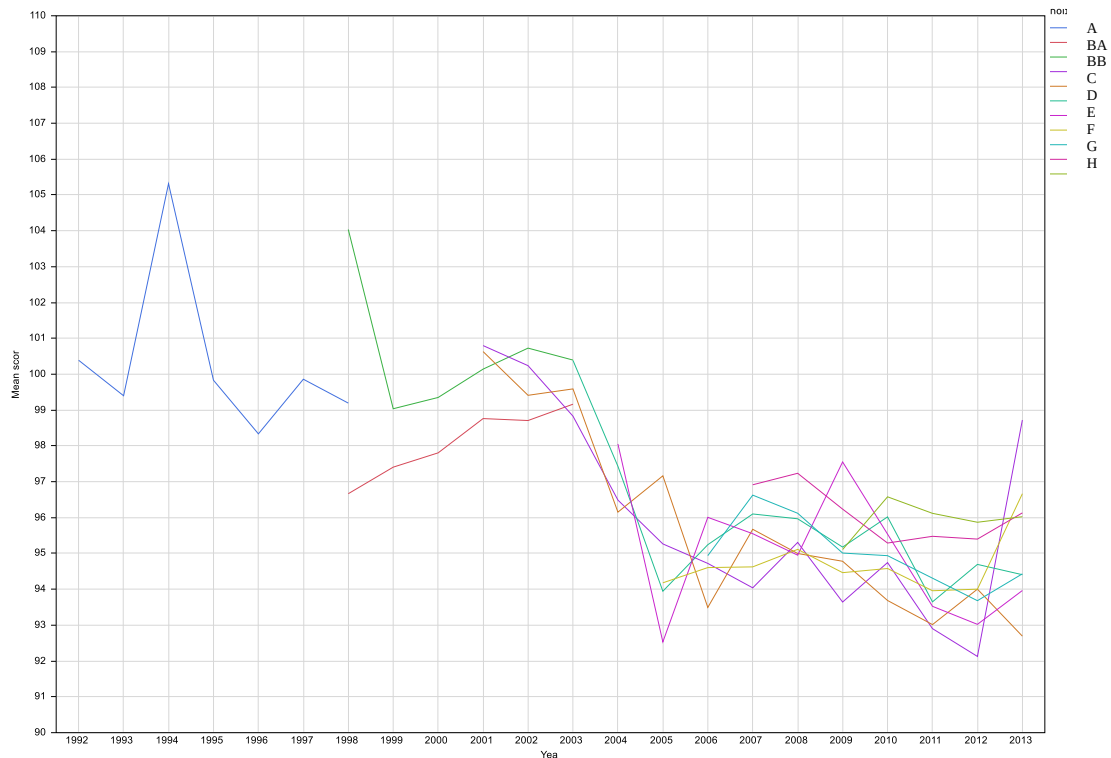
טבלה 6 - שכיחות, ממוצע וסטיית תקן,  
של הציון וה- $\theta$  לפי גרסאות נוסחים

| θ    | ציון  |       |        | N     | גרסת<br>נוסח |
|------|-------|-------|--------|-------|--------------|
|      | ממוצע | ס.ת.  | ממוצע  |       |              |
| 0.97 | -0.02 | 20.44 | 100.13 | 16072 | A            |
| 1.06 | -0.27 | 20.63 | 98.25  | 12996 | BA           |
| 1.07 | -0.18 | 21.00 | 99.85  | 11809 | BB           |
| 1.09 | -0.16 | 22.68 | 99.72  | 4204  | C1           |
| 1.09 | -0.15 | 22.65 | 99.71  | 3970  | D1           |
| 1.15 | -0.10 | 23.41 | 100.39 | 1999  | E1           |
| 1.17 | -0.18 | 24.78 | 97.62  | 3075  | F1           |
| 1.13 | -0.21 | 24.18 | 97.14  | 2914  | E2           |
| 1.14 | -0.33 | 24.69 | 94.93  | 12122 | C2           |
| 1.15 | -0.33 | 24.73 | 94.86  | 12479 | D2           |
| 1.22 | -0.32 | 25.33 | 94.84  | 10310 | F2           |
| 1.19 | -0.32 | 25.35 | 94.72  | 5564  | G2           |
| 1.17 | -0.29 | 24.76 | 95.39  | 15497 | E3           |
| 1.18 | -0.28 | 24.93 | 95.56  | 12116 | H1           |
| 1.20 | -0.24 | 25.40 | 96.45  | 10301 | I1           |
| 1.20 | -0.34 | 25.45 | 94.48  | 9087  | G3           |
| 1.28 | -0.25 | 26.71 | 96.28  | 4869  | J1           |
| 1.26 | -0.37 | 26.36 | 93.61  | 8776  | F3           |
| 1.24 | -0.36 | 26.19 | 94.02  | 2891  | G4           |
| 1.23 | -0.29 | 25.87 | 95.56  | 9848  | I2           |
| 1.27 | -0.27 | 26.57 | 95.88  | 8660  | J2           |
| 1.18 | -0.45 | 25.29 | 92.37  | 1723  | C3           |
| 1.19 | -0.40 | 25.53 | 93.32  | 2389  | D3           |
| 1.24 | -0.36 | 26.01 | 94.01  | 6987  | E4           |
| 1.24 | -0.36 | 26.01 | 94.16  | 7552  | H2           |
| 1.18 | -0.43 | 25.35 | 92.69  | 1436  | D4           |

### תרשימים של ממוצעי הציון בנוסחים השונים לפי עשורים

ניתן לראות בתרשים 8 כי עם התנועה על-פני ציר הזמן משמאל לימין, מספר הנוסחים בכל שנה עולה (נוספים הרבה נוסחי בחינה חדשים ומוצאים מעט נוסחים ישנים); הציון הממוצע בכל שנה הולך ויורד; והתנודתיות (סטיית התקן) בציון הממוצע משנה לשנה הולכת ויורדת אף היא. מן הראוי להזכיר שמספר הנבחנים עולה בכל שנה (ראו תרשים 6).

### תרשים 8 - ציון ממוצע על-פני ציר הזמן לפי נוסחים 1992-2013

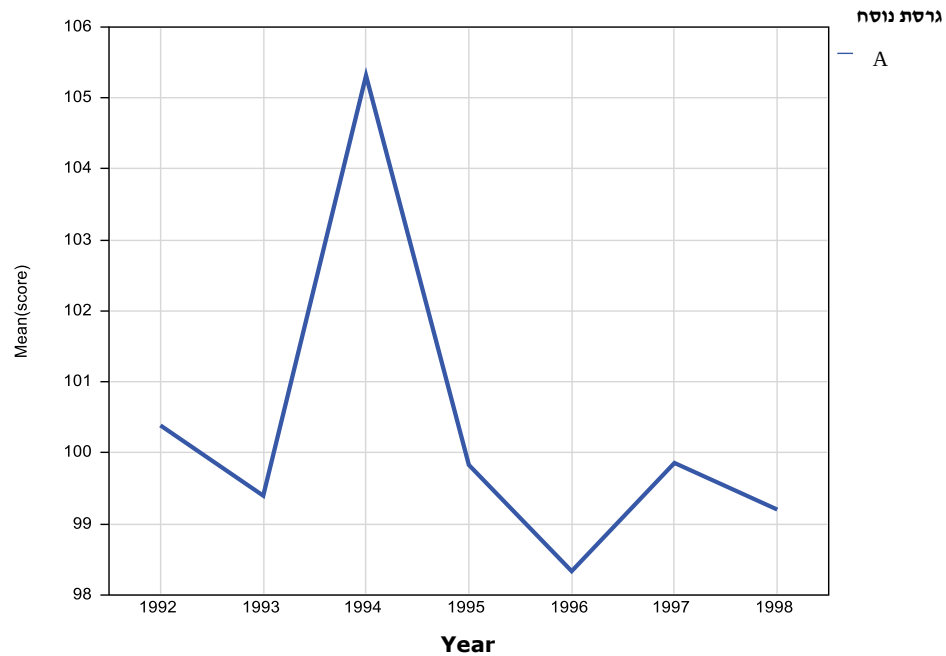


בתרשימים 9 ו-10 מוצגת הסתכלות מפורטת יותר ברמת גרסאות הנוסחים, בחלוקה לפי עשורים. בתרשים 9, בחלק א' של העשור הראשון מ-1992 ועד 1998, הועבר נוסח אמיר"ם אחד בלבד, A. הנוסח התאפיין בתנודתיות רבה מאוד בציון הממוצע משנה לשנה.

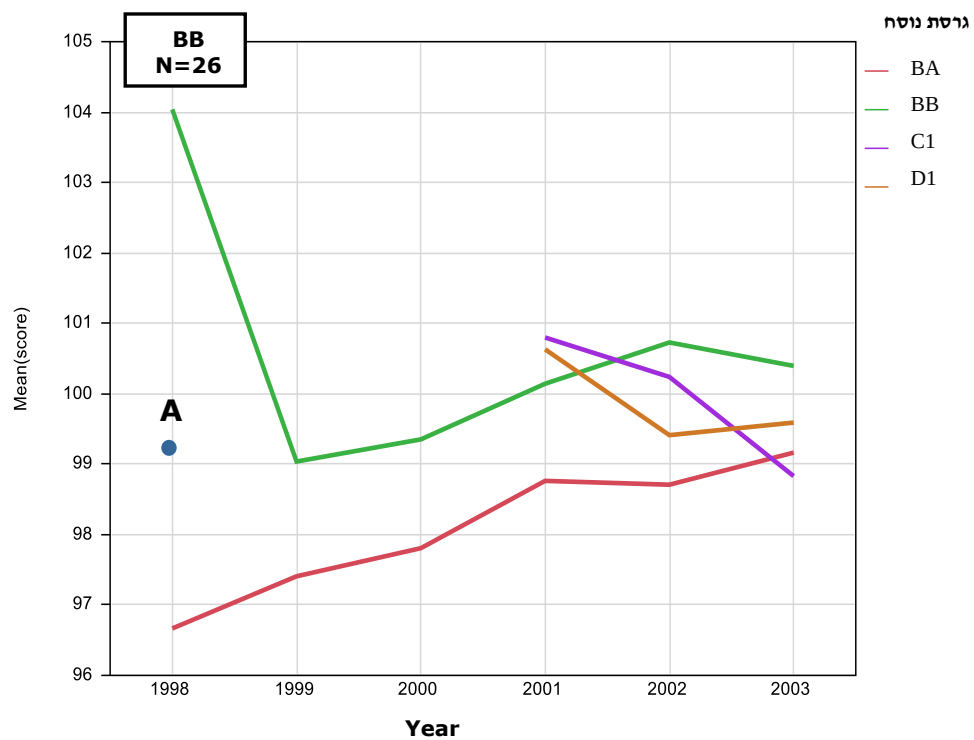
בחלק ב' של העשור הראשון מ-1998 ועד 2003, אפשר לראות כי נוסח A הוצא מהמחזור ב-1998, ובמקומו הוכנסו הנוסחים BA ו-BB. ניתן לראות עליה מונוטונית בציון הממוצע בשני הנוסחים האלו (הציון הממוצע בנוסח BA בשנת 1998 היה גבוה משמעותית באופן בלתי מייצג, בשל מספר הנבחנים הזעום). בשנת 2001 הוכנסו שני נוסחים נוספים, C1 ו-D1, שהראו ציונים דומים מאוד אחד לשני וכן לנוסחים הקודמים. בסך הכול, בעשור הראשון ניכרה תנודתיות רבה של הממוצעים משנה לשנה. לקראת העשור השני, השימוש בכל גרסאות הנוסחים שהוצגו בתרשים 9 הופסק.

תרשים 9 - ציון ממוצע על-פני ציר הזמן לפי גרסת נוסח, 1992-2003

העשור הראשון – חלק א'



העשור הראשון – חלק ב'



בתרשים 10, העשור השני להעברת הבחינה מ-2004 ועד 2013, ניתן לראות שהתקופה מתאפיינת בחלוקה לארבעה מקטעים שונים, כך שעם כל הכנסת קבוצת גרסאות נוסחים חדשה הציון הממוצע משתנה קמעה: המקטע הראשון הוא 2004-2005, השני 2005-2007, השלישי 2007-2011 והרביעי 2011-2013. סך הכול הציון הממוצע מעבר לתקופה די דומה, עם תזוזה של נקודה אחת-שתיים לכאן או לשם.

### תרשים 10 - ציון ממוצע על-פני ציר הזמן לפי גרסת נוסח, 2004-2013



## הבדלים בין נוסחים

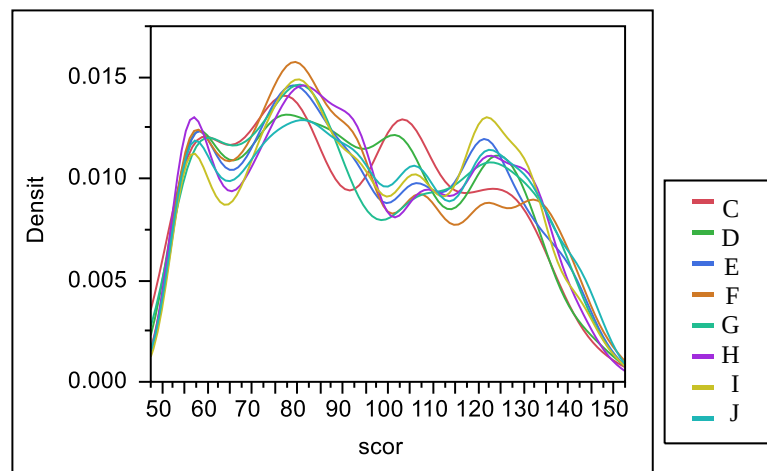
במטרה לבדוק אם קיימים הבדלים בין שמונה הנוסחים שנעשה בהם שימוש בשנים 2012-2013 בוצע ניתוח שונות חד-גורמי ברמת מובהקות של 5% על הציונים. נמצא הבדל מובהק בין נוסחי הבחינה, אך גודל האפקט קטן מאוד,  $F(7, 40377) = 9.29, p < .001, adj. r^2 = .01$ . בטבלה 7 מוצגות תוצאות מבחני הפוסט-הוק מסוג HSD (Honestly Significant Difference של Tukey) על ציון הבחינה הכללי. נוסחים שאין להם אותו סימן (\*, @, #) שונים במובהק זה מזה. מהטבלה עולה כי אף נוסח אינו שונה במובהק מכל יתר הנוסחים גם יחד (אך יש נוסחים מסוימים השונים מאחרים).

טבלה 7 - פוסט-הוק מסוג HSD על ציון הבחינה הכללי 2012-2013

| נוסח |   | ממוצע | N    |
|------|---|-------|------|
| J    | * | 95.96 | 7264 |
| I    | * | 95.82 | 8317 |
| E    | @ | 94.53 | 5658 |
| H    | # | 94.16 | 7552 |
| G    | @ | 94.05 | 1623 |
| F    | # | 93.53 | 6347 |
| D    | # | 93.30 | 2685 |
| C    | # | 92.25 | 939  |

תרשים 11 מציג את צפיפות הציונים בכל רמת ציון, וניתן ללמוד ממנו כי למרבית הנוסחים התפלגות דומה.

תרשים 11 - צפיפות: התפלגות הציון לפי נוסח 2012-2013



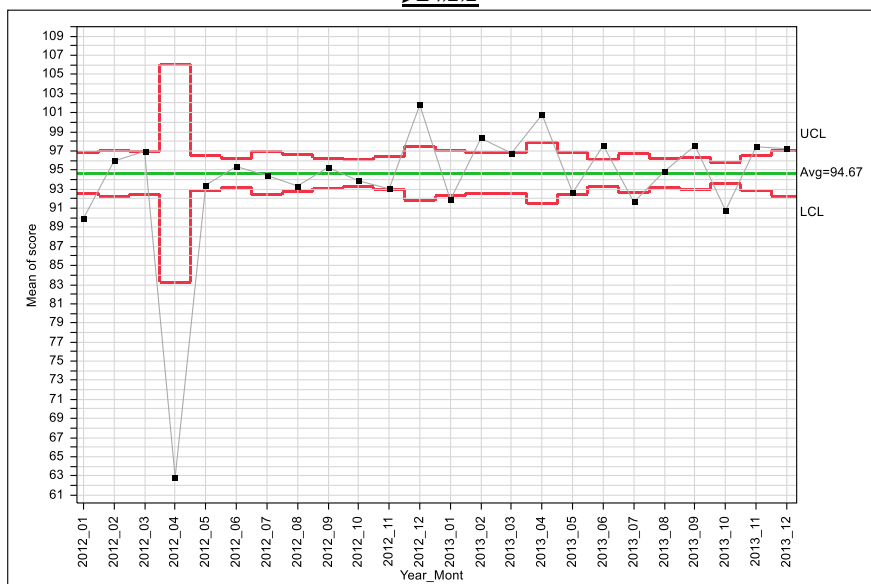
## תרשימי בקרה

תרשים בקרת איכות מציג את ההיסטוריה המצטברת של מדד סטטיסטי. התרשים משורטט על-פני שני ממדים, זמן - על הציר האופקי ותצפית - על הציר האנכי, וכולל שלושה קווים: קו אמצעי – קו הבסיס (ממוצע), קו גבול עליון וקו גבול תחתון במרחק שווה מעל ומתחת לקו האמצעי (3 סטיות תקן מעל ומתחת לקו הבסיס, בהתאמה). נקודה שנופלת בין שני קווי הגבול ודפוס סטייה רנדומאלי מעידים שהתהליך "בשליטה" (להרחבה על אודות השיטה ר' Shewhart, 1931, 1939).

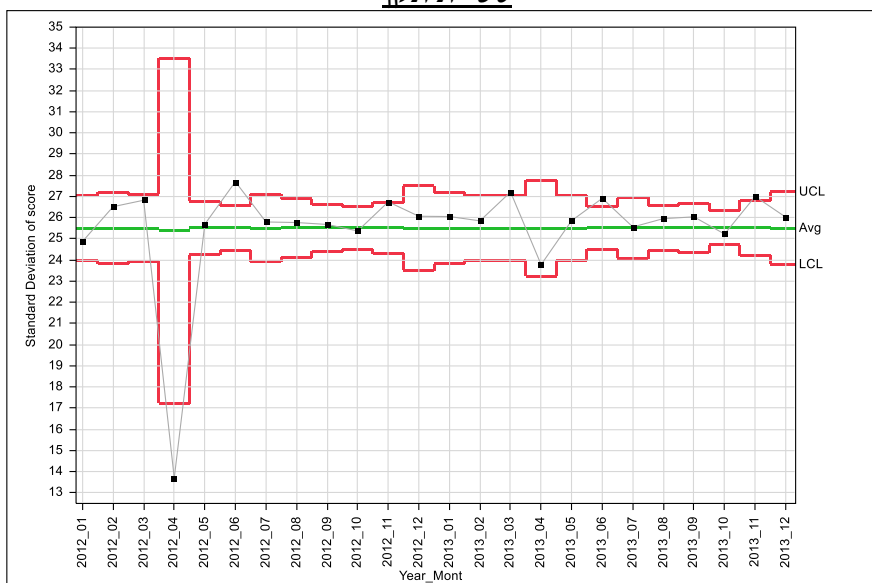
בקרת האיכות נערכה על התקופה 2012-2013. תרשים 12 מציג תרשימי בקרה (ממוצע וסטיית תקן) על הציונים הגולמיים. בעוד שסטיות התקן יציבות למדי, בממוצעים נמצאו חריגות. החריגה הבולטת ביותר היא באפריל 2012. כאשר בודקים חריגה זו מסתבר כי הנתון מתבסס על מספר קטן יחסית של נבחנים (45 איש) ממוסד אקדמי ספציפי.

### תרשים 12 - תרשימי בקרה על הציונים 2012-2013

#### ממוצע



#### סטיות תקן

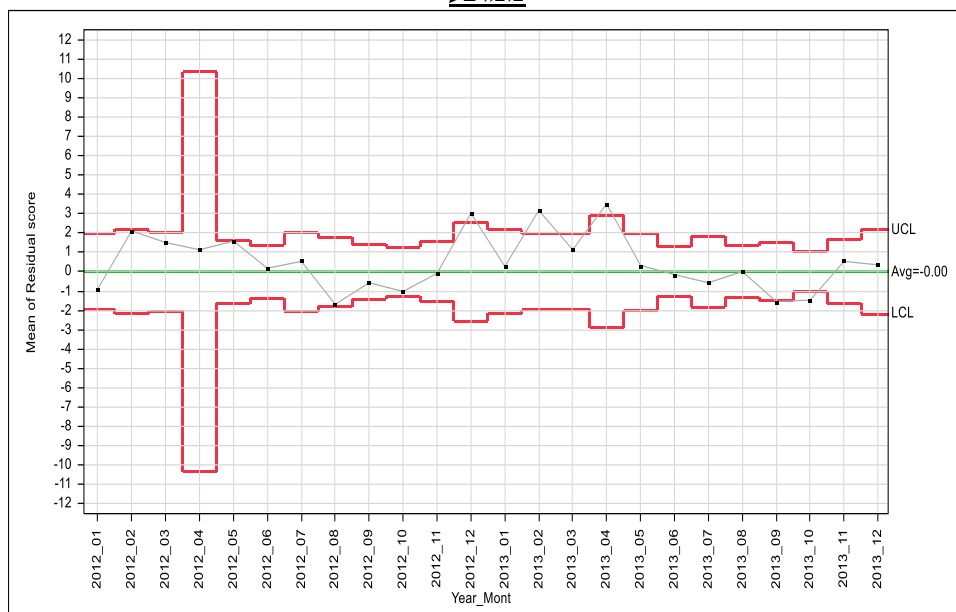


כעת נבדק האם ה"שונות הלא מוסברת" (שונות שלא ניתן להסביר ע"י משתנים אחרים) היא "בשליטה". לצורך כך, חושבה רגרסיה לניבוי ציון הבחינה באמצעות המוסד אקדמי בו נבחנו המשתתפים. הניבוי נמצא מובהק,  $F(49, 40335) = 192.32, r^2 = 0.19, Adj. r^2 = 0.19, p < .001$ .

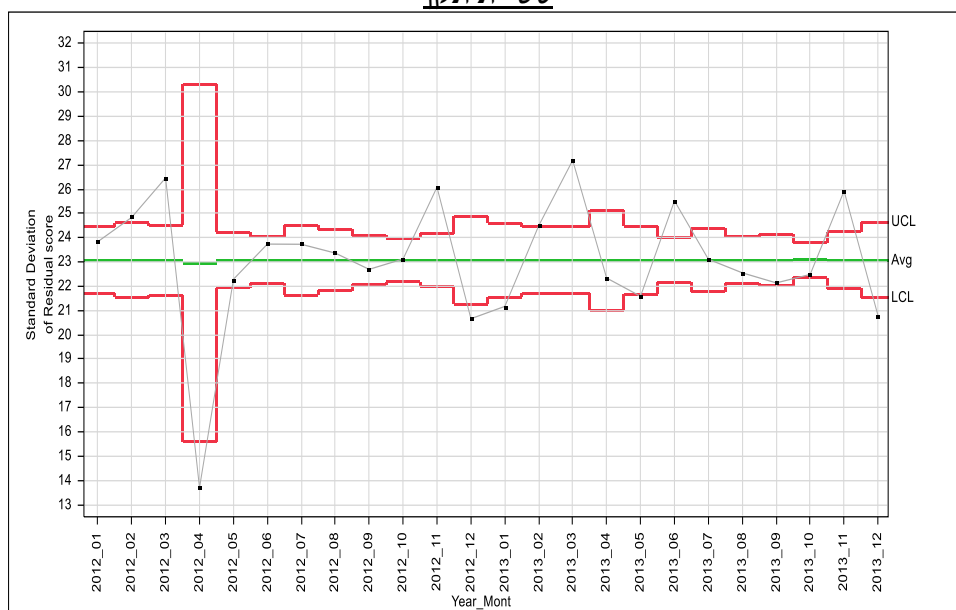
לכל נבחן חושבה השארית (הפרש בין הציון האמיתי לציון המנובא). בתרשים מספר 13 מוצגים תרשימי הבקרה (ממוצע וסטיית תקן) של שאריות הציונים. בהסתכלות על ממוצעי השאריות ניתן לראות מעט מאוד חריגות. אפילו החריגה הבולטת ביותר של אפריל 2012 נעלמה, דבר המעיד על היותה תולדה של הבדלים בין המוסדות האקדמיים (שמושכים סטודנטים עם הישגים דומים) ולא כפועל יוצא מטעות או בעיה במבחן. כמו כן רואים כי סטיות התקן איבדו מיציבותן (לכן יש לקחת ממצאים אלו בעירבון מוגבל).

### תרשים 13 - תרשימי בקרה על שאריות הציונים 2012-2013

#### ממוצע



#### סטיית תקן



## ניתוח הציונים של נבחנים חוזרים

באמיר"ם היו 199,674 היבחנויות בין השנים 1992 (אוגוסט) עד 2013 (דצמבר). מתוכן 152,952 היו בחינות ראשונות, והיתר (23.4%) בחינות חוזרות, כלומר, הועברו לנבחנים שכבר נבחנו בעבר בבחינת אמיר"ם במהלך שנים אלה (מספר הבחינות החוזרות נע בין 2 ל-12). מאחר שמספר הנבחנים הניגשים ליותר מארבע בחינות נמוך מאוד (0.4% מהנבחנים) לא הושו נתוני העברות אלו ליתר ההעברות.

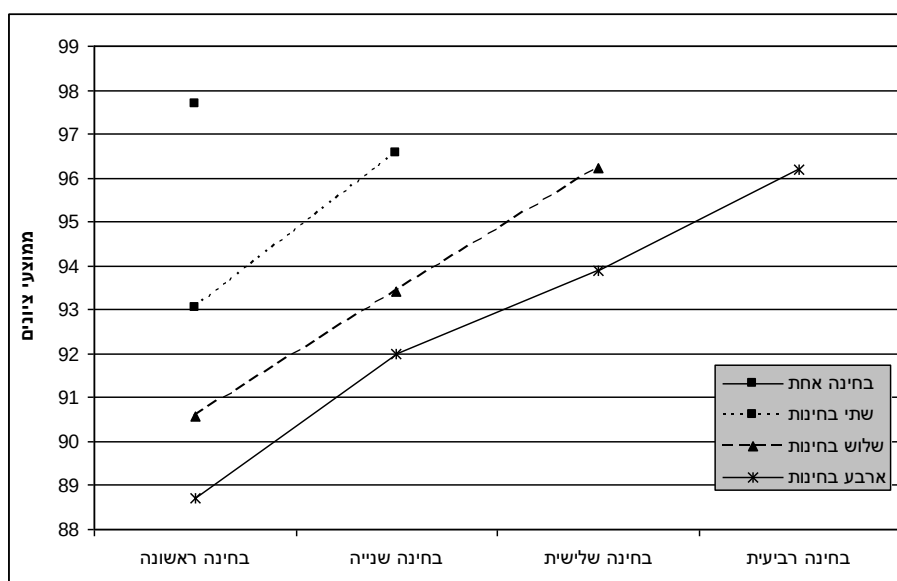
בטבלה 8 ובתרשים 14 מוצגת התפלגות הציונים של נבחנים לפי מספר הבחינות שהם ביצעו (סטטיסטיים בין-נבחניים). למשל, נבחן שנבחן רק פעם אחת, נתוניו יופיעו בשורת "בחינה אחת" בלבד; נבחן שנבחן פעמיים, נתוני בחינתו השנייה יופיעו בשורת "שתי בחינות" בלבד; נבחן שנבחן שלוש פעמים נתוני בחינתו השלישית יופיעו בשורת "שלוש בחינות" בלבד, וכן הלאה. ניתן לראות שעבור כל קבוצת נבחנים, ככל שמספר החזרות על הבחינה עולה – ממוצע הציון הולך ועולה.

**טבלה 8 - ממוצע (ס.ת.) ומספר נבחנים בקבוצת נבחנים לפי מספר בחינה 1992-2013**

| קבוצת נבחנים       | בחינה ראשונה     | בחינה שנייה      | בחינה שלישית     | בחינה רביעית     | N       |
|--------------------|------------------|------------------|------------------|------------------|---------|
| <b>בחינה אחת</b>   | 97.68<br>(25.86) |                  |                  |                  | 117,231 |
| <b>שתי בחינות</b>  | 93.04<br>(21.75) | 96.55<br>(22.87) |                  |                  | 27,591  |
| <b>שלוש בחינות</b> | 90.57<br>(21.34) | 93.4<br>(21.62)  | 96.23<br>(22.47) |                  | 6,112   |
| <b>ארבע בחינות</b> | 88.70<br>(20.60) | 91.99<br>(21.4)  | 93.88<br>(21.3)  | 96.20<br>(22.01) | 1,450   |

\* נתוניהם של 568 נבחנים שנבחנו 5 פעמים ומעלה לא מופיעים בטבלה

**תרשים 14 - ממוצעי הציונים בקרב הנבחנים החוזרים**





## נתונים תוך-נבחניים

בטבלה 9 מוצגים סטטיסטיים תוך-נבחניים. למשל, נבחן שנבחן שלוש פעמים בתקופה המדוברת, נתוני בחינתו הראשונה ישתקללו בשורת ההעברה הראשונה, נתוני בחינתו השנייה ישתקללו בשורת ההעברה השנייה, ונתוני בחינתו השלישית ישתקללו בשורת ההעברה השלישית. ניתן לראות כי הציונים דומים מאוד.

**טבלה 9 - ממוצע, סטיית תקן ומספר נבחנים לפי מספר היבחנות 1992-2013**

| נבחנו לפחות | ממוצע | סטיית תקן | N       |
|-------------|-------|-----------|---------|
| פעם אחת     | 96.42 | 25.04     | 152,952 |
| שתי פעמים   | 95.69 | 22.64     | 35,721  |
| שלוש פעמים  | 95.32 | 22.27     | 8,130   |
| ארבע פעמים  | 94.72 | 21.93     | 2,018   |

\* נתונים של 853 נבחנים שנבחנו 5 פעמים ומעלה לא מופיעים בטבלה

על-מנת לאמוד את מידת השיפור בהישגים מבחינה לבחינה, נבדק במערך תוך-נבחני הפרש הציונים בקרב קבוצת הנבחנים החוזרים בלבד, דהיינו – מי שנבחנו פעמיים עד ארבע פעמים, 35,709 נבחנים ייחודיים. בטבלה 10 מוצגת ההתפלגות של הפרשי הציון.

**טבלה 10 - ממוצע וסטיית תקן של הפרשי הציונים והפרש מתוקנן בין ההיבחנות ומספר הנבחנים 1992-2013**

| הפרשי ציון הבחינה הכללי | ממוצע ההפרש | סטיית תקן של ההפרש | הפרש מתוקנן <sup>2</sup> | N      |
|-------------------------|-------------|--------------------|--------------------------|--------|
| בין העברה 2 להעברה 1    | 3.39        | 10.06              | 0.14                     | 35,721 |
| בין העברה 3 להעברה 1    | 5.49        | 10.95              | 0.22                     |        |
| בין העברה 3 להעברה 2    | 2.58        | 10.46              | 0.11                     | 8,130  |
| בין העברה 4 להעברה 1    | 7.07        | 11.11              | 0.29                     |        |
| בין העברה 4 להעברה 2    | 3.90        | 10.77              | 0.16                     |        |
| בין העברה 4 להעברה 3    | 2.28        | 10.72              | 0.09                     | 2,018  |

כפי שניתן לראות בטבלה, כשמתבוננים על העברות עוקבות (בין העברה 2 ל-1, בין העברה 3 ל-2 ובין העברה 4 ל-3) – השיפור הממוצע בין ציוני ההעברות יורד ככל שמספר ההיבחנות עולה, ואילו סטיית התקן עולה במעט ככל שמספר ההיבחנות עולה.

<sup>2</sup> מתוקנן לפי סטיית התקן של הבחינה 24.49.

## הפרש הזמן בין בחינות חוזרות

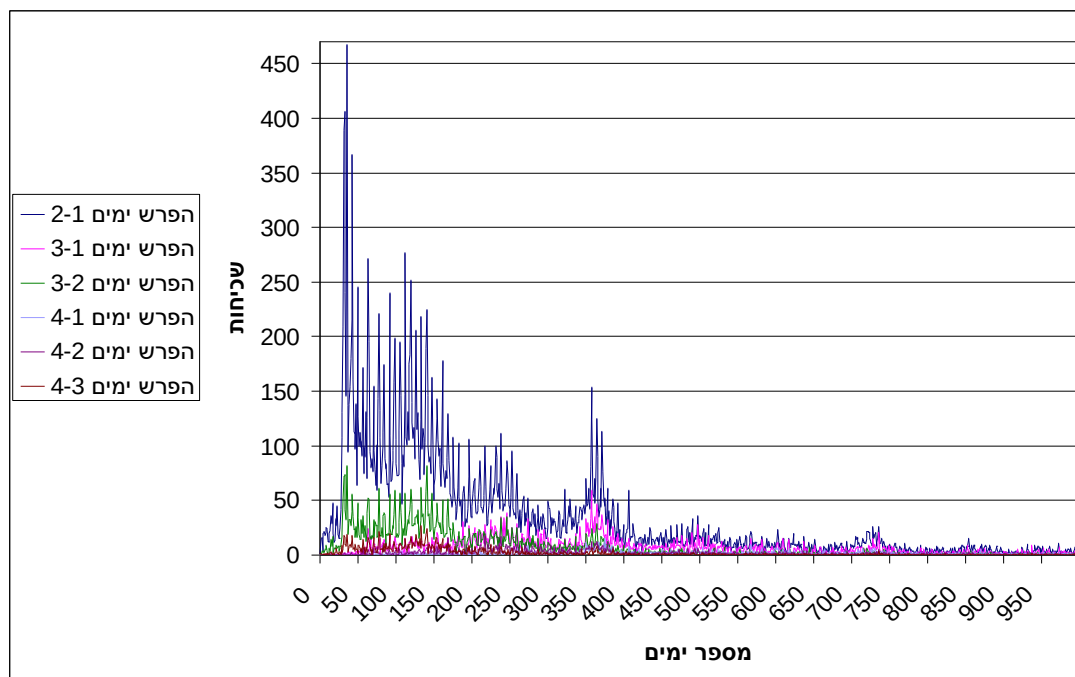
בטבלה 11 מוצגים סטטיסטיים אודות מרווחי הזמן בין ההיבחנויות בימים. ניתן לראות שבין העברות עוקבות (בין העברה 2 ל-1, בין העברה 3 ל-2 ובין העברה 4 ל-3) יש הפרש ימים חציוני בסדר גודל של חצי שנה, עם שונות גדולה.

**טבלה 11 - מספר הנבחנים, ממוצע, סטיית תקן וחציון של מרווחי הזמן (בימים) בין היבחנויות 1992-2013**

| הפרשי ימים           | ממוצע  | סטיית תקן | חציון | N      |
|----------------------|--------|-----------|-------|--------|
| בין העברה 2 להעברה 1 | 306.55 | 452.50    | 162   | 35,721 |
| בין העברה 3 להעברה 1 | 545.41 | 564.10    | 369   |        |
| בין העברה 3 להעברה 2 | 253.05 | 340.63    | 154   | 8,130  |
| בין העברה 4 להעברה 1 | 753.10 | 655.46    | 562   |        |
| בין העברה 4 להעברה 2 | 457.23 | 462.26    | 329   |        |
| בין העברה 4 להעברה 3 | 225.53 | 317.69    | 140   | 2,018  |

טווחי הזמן בין ההיבחנויות נעים עד כ-19 שנים. תרשים 15 מציג הפרשי ימים מ-0 ועד 2000 יום – טווח המכסה מעל ל-95 אחוז מהתצפיות. ניתן לראות כי מדובר בהתפלגות א-סימטרית עם זנב ימני ארוך. כלומר, רוב הנבחנים נבחנים בתום תקופה קצרה לאחר בחינתם הראשונה, ובחלוף הזמן מספר הנבחנים יורד משמעותית.

**תרשים 15 - התפלגות הפרשי זמני ההעברות בימים 1992-2013**



### מעקב אחר נוסחי הבחינה שנבחנו חוזרים מקבלים

בטבלה 12 מוצג מספר הפעמים שנבחן קיבל את אותו נוסח (מאגר של מאות פריטים פריטים), לפי נוסח. למשל, נוסח A התקבל פעם אחת בלבד עבור 1108 נבחנים, התקבל פעמיים בלבד עבור 1899 נבחנים, התקבל שלוש פעמים בלבד עבור 381 נבחנים וכן הלאה. הנוסחים מסודרים לפי סדר כרונולוגי (טורים – מימין לשמאל); ככל שהנוסח ישן יותר, מספר הנוסחים שהועברו במקביל מין הסתם קטן יותר.

טבלה 12 - מספר הנבחנים שקיבלו כל נוסח  
לפי נוסח ומספר הפעמים שהנוסח התקבל 1992-2013

| ספר פעמים<br>הנוסח התקבל | A     | BA    | BB    | C     | D     | E     | F     | G     | H     | I     | J     |
|--------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| פעם אחת                  | 1,108 | 3,609 | 3,364 | 6,279 | 6,956 | 8,361 | 6,968 | 5,844 | 6,129 | 5,804 | 3,874 |
| דיוק פעמיים              | 1,899 | 1,207 | 1,020 | 692   | 653   | 1,189 | 889   | 718   | 741   | 827   | 467   |
| דיוק שלוש פעמים          | 381   | 133   | 101   | 36    | 30    | 71    | 42    | 35    | 28    | 33    | 19    |
| דיוק ארבע פעמים          | 91    | 18    | 10    | 4     | 1     | 5     | 1     | 1     | 2     | 1     | 0     |
| דיוק חמש פעמים           | 29    | 5     | 1     | 0     | 0     | 1     | 1     | 0     | 0     | 0     | 0     |
| דיוק שש פעמים            | 6     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 1     | 0     | 0     |
| דיוק שבע פעמים           | 3     | 0     | 0     | 1     | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
| דיוק שמונה פעמים         | 1     | 0     | 0     | 0     | 1     | 0     | 0     | 0     | 0     | 0     | 0     |

בטבלה 13 מוצגת בדיקה סטטיסטית של השיפור בציון בקרב נבחנים שקיבלו נוסח זהה ושונה בשתי העברות. ניתן לראות, כי הסיכוי לקבל את אותו נוסח בשתי העברות עומד על קצת יותר מ-20 אחוזים. בנוסף, ההפרש בין ממוצעי הציונים גבוה (מובהק ומובהק שולית) לטובת מי שקיבל את אותו נוסח בשתי העברות עוקבות אך לא כך בהעברות שאינן עוקבות. כשבוחנים את השיפור בפרספקטיבה של הבחינה כולה – במונחי סטיית תקן - גודלו זעיר.

טבלה 13 - שיעור הנבחנים שקיבלו נוסח זהה או שונה בהעברה חוזרת, ממוצע הפרשי הציונים (ס.ת.), הפרש מתוקנן, בחינת מובהקות ושכיחות לפי מספר העברה 1992-2013

| נוסח זהה     | נוסח שונה    | נוסח זהה     | נוסח שונה    | הפרש מתוקנן <sup>3</sup> | תוצאה |
|--------------|--------------|--------------|--------------|--------------------------|-------|
| %            | %            | %            | %            |                          |       |
| ממוצע (ס.ת.) | ממוצע (ס.ת.) | ממוצע (ס.ת.) | ממוצע (ס.ת.) |                          |       |
| N            | p            | t            |              |                          |       |
| העברה 1      | 23.5         | 76.5         | 3.79 (0.11)  | 3.26 (0.06)              | 0.02  |
| העברה 1      | 21.8         | 78.2         | 5.37 (0.26)  | 5.52 (0.14)              | -0.01 |
| העברה 2      | 24.6         | 75.4         | 3.00 (0.23)  | 2.44 (0.13)              | 0.02  |
| העברה 1      | 20.4         | 79.6         | 6.85 (0.55)  | 7.13 (0.28)              | -0.01 |
| העברה 2      | 23.0         | 77.0         | 4.27 (0.50)  | 3.78 (0.27)              | 0.02  |
| העברה 3      | 24.7         | 75.3         | 3.08 (0.48)  | 2.02 (0.27)              | 0.04  |

\* מובהק

\*\* מובהק שולית

<sup>3</sup> מתוקנן לפי סטיית התקן של הבחינה 24.49.

בקרת איכות חיונית להבטיח את מהימנות המבחן, החישוב והדיווח של הציונים לנבחנים. לבקרה לאורך זמן של תהליך הציון – במיוחד כשמדובר בבחינה המועברת שנים ארוכות – חשיבות רבה, הן בהסתכלות על העבר והן בפיתוח שיטות ושגרת בקרות לבחינת הציונים בהווה. המחקר הנוכחי תרם להרחבת הידע בנוגע לבקרת איכות על מבחנים המועברים תכופות לקבוצות נבחנים קטנות, תוך נתינת מענה ראשוני גם למבחנים אדפטיביים בנוסף למבחנים לינאריים.

בסיס הנתונים של המרכז הארצי לבחינות ולהערכה עמד לרשות החוקרים. בין השנים 1992 עד 2013 התקיימו באמיר"ם היו – 199,674 היבחנויות ב-11 נוסחים. התכנסות הבחינה עומדת על כ-80 אחוז בממוצע. ציון  $\theta$  (אינדיקטור ליכולת) וציון הבחינה, המתקבל על ידי טבלת המרה, נמצאים בקשר כמעט מושלם, והופכים להיות מתואמים יותר ויותר עם הזמן.

### סיכום 20 שנות העברה של בחינת אמיר"ם

לצד העלייה העקבית במספר הנבחנים לאורך השנים, ניכרת ירידה עקבית בציון הממוצע בבחינה. גיל הנבחנים הממוצע הינו כ-27 (ס.ת. 8), וגברים (43%) מגיעים בממוצע להישגים גבוהים משל נשים (בכ-0.4 ס.ת.). ישנו שימוש גובר בבחינה במוסדות חדשים לצד הוותיקים. מספר הנבחנים באמיר"ם עולה כיום על מספר הנבחנים באמיר"ר, ואילו הציון הממוצע באמיר"ם נמוך מזה של אמיר"ר.

### בקרת איכות

בהתייחס לשלושת הרכיבים של בקרת האיכות נמצא כי –

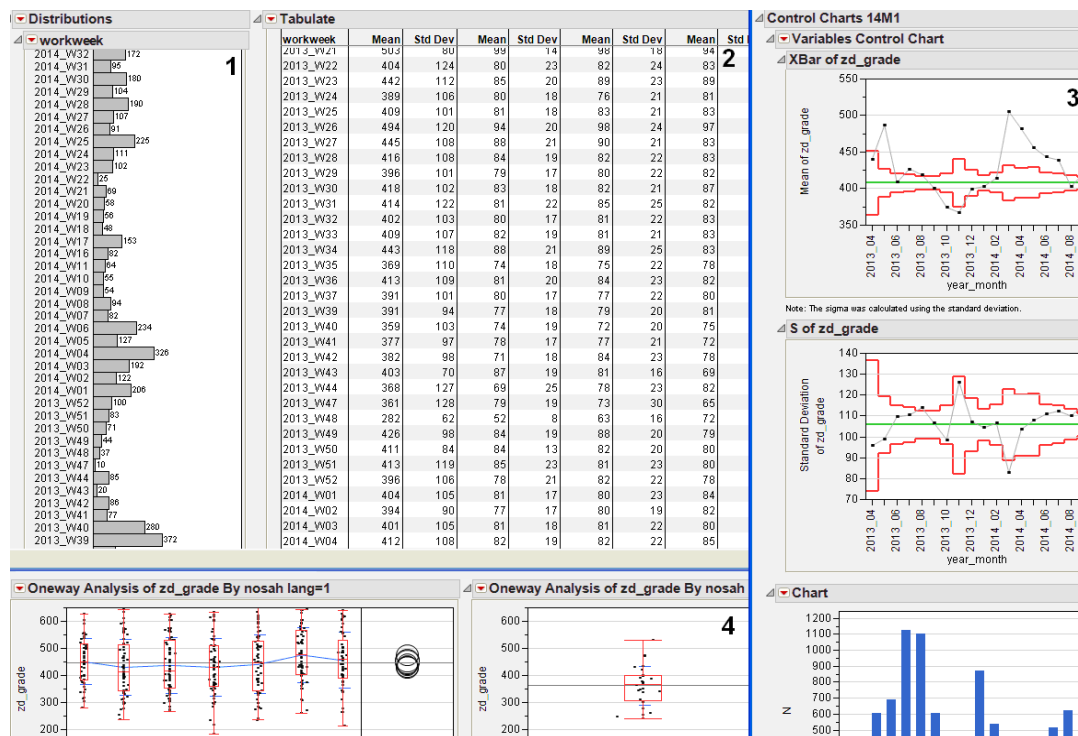
**נוסח.** הציון הממוצע וסטיות התקן של הנוסחים השונים יציבים למדי, במיוחד בעשור האחרון. נמצא הבדל מובהק אך לא משמעותי בממוצעי הנוסחים 2012-2013.

**תרשימי בקרה.** בקרת איכות על ציוני הבחינה, תוך ניכוי השפעת המוסד על הציון, מראה כי תהליך הציון הינו יחסית "בשליטה" (נסייג ממצאים אלו לאור התנודות בסטיות התקן).

**הישגי נבחנים חוזרים.** כרבע מהבחינות בבסיס הנתונים היו בחינות חוזרות. הפרש הימים החציוני בין העברות עוקבות עומד על כחצי שנה, ובחלוף הזמן מספר הנבחנים החוזרים יורד משמעותית. ככל שמספר החזרות על הבחינה עולה – ממוצע ציוני הנבחנים החוזרים הולך ועולה. נמצא שככל שתאריך הבחינה מאוחר יותר, קטן הסיכוי לחזרה על אותו הנוסח. גם קבלת אותו נוסח אינה מניבה ציון גבוה משמעותית מחזרה על הבחינה בנוסח מקביל: נמצא שיפור מובהק אך קטן לטובת אלו שקיבלו את אותו נוסח בשתי העברות עוקבות בלבד, אך לא בהעברות שאינן עוקבות. נדגיש, כי הגם שמדובר באותו נוסח לא מדובר באותם פריטים בדיוק שכן הסיכוי לדגימת אותם פריטים ממאגר הפריטים הינו נמוך. עם זאת, מאחר ויש אפקט, למרות שהוא קטן, יש למצוא דרכים להקטינו עוד יותר או אף למנוע חזרה על אותו נוסח.

אחד מתוצרי המחקר הינו מערכת ממוכנת אוטומטית לבקרת ציונים שבועית טרם דיווחם (באומר, בהכנה). תרשים 16 מציג דוגמא הלקוחה ממערכת דומה שכבר פועלת על בחינה אחרת, בחינת המימ"ד. בתרשים, מוצגת הסטטיסטיקה התיאורית (מסומן כ-1), טבלאות נתונים (2), טבלאות בקרה של שוהארט (3) וכן בחינת מובהקות סטטיסטית (4).

תרשים 16 - צילום מסך מערכת ממוכנת לבקרה מתמשכת על בחינת המימ"ד 2014



## מחקרי המשך

מחקר עתידי יתמקד בהחלת שגרת בקרת האיכות שפותחה במחקר הנוכחי גם על בחינות נוספות. בנוסף, מחקר עתידי ימשיך ויבחן ביצועי נבחנים חוזרים בבחינות ידע באנגלית תוך השוואה בין בחינות שונות (אמיר"ם, אמיר"ר וחלק האנגלית של הבחינה הפסיכומטרית), אדפטיביות וליניאריות כאחד. באופן ספציפי, ניתן לבחון האפקט של סדר ההיבחנות על השיפור (קודם באמיר"ר ואז באמיר"ם ואז להפך, לדוגמא). שאלה מעניינת נוספת היא מהו גודל ההבדל בבחינות מקבילות בין חזרה על נוסח בבחינה ליניארית בה כל הפריטים זהים לעומת אדפטיבית בה חלק מהפריטים זהים. בנוסף, ניתן לבחון מהו אחוז הפריטים החופף אצל נבחנים החוזרים על נוסח ספציפי בבחינת אמיר"ם, מהו סוג הפריטים החוזר, והאם לדבר השפעה על הציון. בקורות נוספות יתמקדו ברמת הפריטים, באופי האדפטיבי של הבחינה ובמהלכה בהתאם למודל התיאורטי.

אללוף, א', גוטנטג, ט', באומר, מ', לוי, ס', פרונטון, מ' ורודד, ק' (2014). *פיתוח מדדי לבקרת איכות לציוני מבחנים המועברים תכופות*. דוח מספר 400, מרכז ארצי לבחינות ולהערכה. ISBN: 978-965-502-181-3

באומר, מ' (בהכנה). *מערכת בקרת איכות למבחנים ממוחשבים*. דוח טכני, מרכז ארצי לבחינות ולהערכה.

באומר, מ', אללוף, א', גוטנטג, ט' ופרונטון, מ' (2014). *פיתוח מדדי בקרת איכות לציוני מבחנים המועברים תכופות*. הכנס ה-10 של האגודה הישראלית לפסיכומטריקה (אפ"י), רמת גן, ישראל.

גוטנטג, ט', אללוף, א', באומר, מ', פרונטון, מ' ופודולר, נ' (2015). *21 שנות בחינת אמיר"ם – בקרת איכות ומחקר*. הכנס ה-11 של האגודה הישראלית לפסיכומטריקה (אפ"י), ירושלים, ישראל.

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*, 261 (1), 36-46.

Allalouf, A. & Gutentag, T. (under revision). Quality Control for Scoring Tests Administered in Continuous Mode. *Educational Measurement: Issues and Practice*.

Allalouf, A., & Gutentag, T. (April, 2015). *Developing Quality Control Procedures for Continuously Administered Tests*. Eposter presented at the National Council on Measurement in Education, Chicago, Illinois.

Allalouf, A., Gutentag, T., Baumer, M., & Fronton, M. (July, 2014). *Quality Control for Scoring Continuously Administered Tests*. Presented at the 9<sup>th</sup> Conference of the International Test Commission, San Sebastian, Spain.

Gutentag, T., Allalouf, A., Baumer, M., Levy, S., Fronton, M., & Roded, K. (October, 2013). *Quality Control for Scoring Continuously Administered Tests*. Presented at the 39<sup>th</sup> Annual International Association for Educational Assessment Conference, Tel Aviv, Israel.

ITC (International Test Commission) (2014). ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. *International Journal of Testing*, 3, 195-217.

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. *Psychometrika*, 78 (3), 557-575.

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), 1-7.

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.