21 שנות בחינת אמיר"ם בקרת איכות ומחקר

טוני גוטנטג, אבי אללוף, מיכל באומר, מרינה פרונטון ונדב פודולר
מרכז ארצי לבחינות ולהערכה

כנס 2015 של האגודה הישראלית לפסיכומטריקה (אפ"י)

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

שני סוגי בקרת איכות

סוג ראשון

בקרת איכות לבחינות המועברות לקבוצות אוכלוסייה גדולות במספר קטן של מועדים קבועים תוך שימוש במספר נוסחים קטן.

למשל, הפסיכומטרי, SAT, ACT

סוג שני

בקרת איכות לבחינות המועברות לקבוצות אוכלוסייה קטנות במספר תאריכים רב תוך שימוש במספר רב של נוסחי בחינה.

למשל, אמיר"ם, מימ"ד, IELTS, TOEFL

תאריך	מספר נוסחים	מספר נבחנים	
		סה"כ	ממוצע לנוסח
5 באפריל	1	3000	3000
7 ביוני	2	9000	4500
11 ביולי	2	8000	4000

תאריך	מספר נוסחים	מספר נבחנים	
		סה"כ	ממוצע לנוסח
5 באפריל	6	120	20
7 באפריל	7	210	30
11 באפריל	6	90	15

בקרת איכות על שני סוגי המבחנים

- לסוג הבחינות הראשון והמסורתי פותחו שיטות בקרת איכות ידועות ומוכרות (Allalouf, 2007; Kolen & Brennan, 2004).
- שיטות אלה אינן שמישות בבחינות מהסוג השני, בעיקר בשל גודלי המדגמים הקטנים ובשל מספר רב של נוסחים המועברים במקביל.

- מחקר זה עוסק **בסוג השני**, שהינו סוג רלוונטי במיוחד של בקרת איכות, בהינתן שעולם הבחינות מתקדם בקצב מהיר לכיוון זה.
- ניתן לראות בבקרת איכות מתמשכת כעין "**רפואה מונעת**" תקלות עתידיות.

מטרת המחקר הנוכחי

- בקרת איכות על ציוני בחינה היא הכרחית עבור הגופים העורכים מבחנים, הגופים המשתמשים בציונים והנבחנים,
במיוחד כשמדובר במבחנים **עתירי סיכון** (high stake).

1. התבוננות "**היסטורית**" ארוכת טווח על הציונים ואיכותם,
שכן בחינת אמיר"ם מועברת מזה שני עשורים.

2. עריכת **בקרת איכות** על בחינת אמיר"ם
המועברת באופנות העברה מתמשכת,
בהסתמך על שיטות סטטיסטיות שנחקרו ופותחו לאחרונה.
(Lee & von Davier, 2013 ;Schafer et al., 2011)

שיטה

שיטת המחקר

- **אוכלוסייה**
  - 199,674 נבחנים
  - בין השנים 1992 (אוגוסט) עד 2013 (דצמבר)
  - ב-27 נוסחים (מתוכם ניתן לקבץ 10 נוסחים)

מבחני המרכז הארצי
הבוחנים רמת אנגלית 2014

	הבחינה הפסיכומטרית	בחינת מימ"ד	אמי"ר	אמיר"ם
מטרה	כלי לחיזוי הצלחה אקדמית; מיון לחוגים; מיון רמת אנגלית.	בחינת ידע למועמדים למכינות הקדם-אקדמיות; מיון למסלולים.	מיון רמת אנגלית במוסדות לימוד אקדמיים.	
תכנים	מילולי, כמותי, אנגלית	עברית, מתמטיקה, אנגלית	אנגלית	
אופנות	נייר ועיפרון	ממוחשב	נייר ועפרון	ממוחשב אדפטיבי

בחינת אמיר"ם

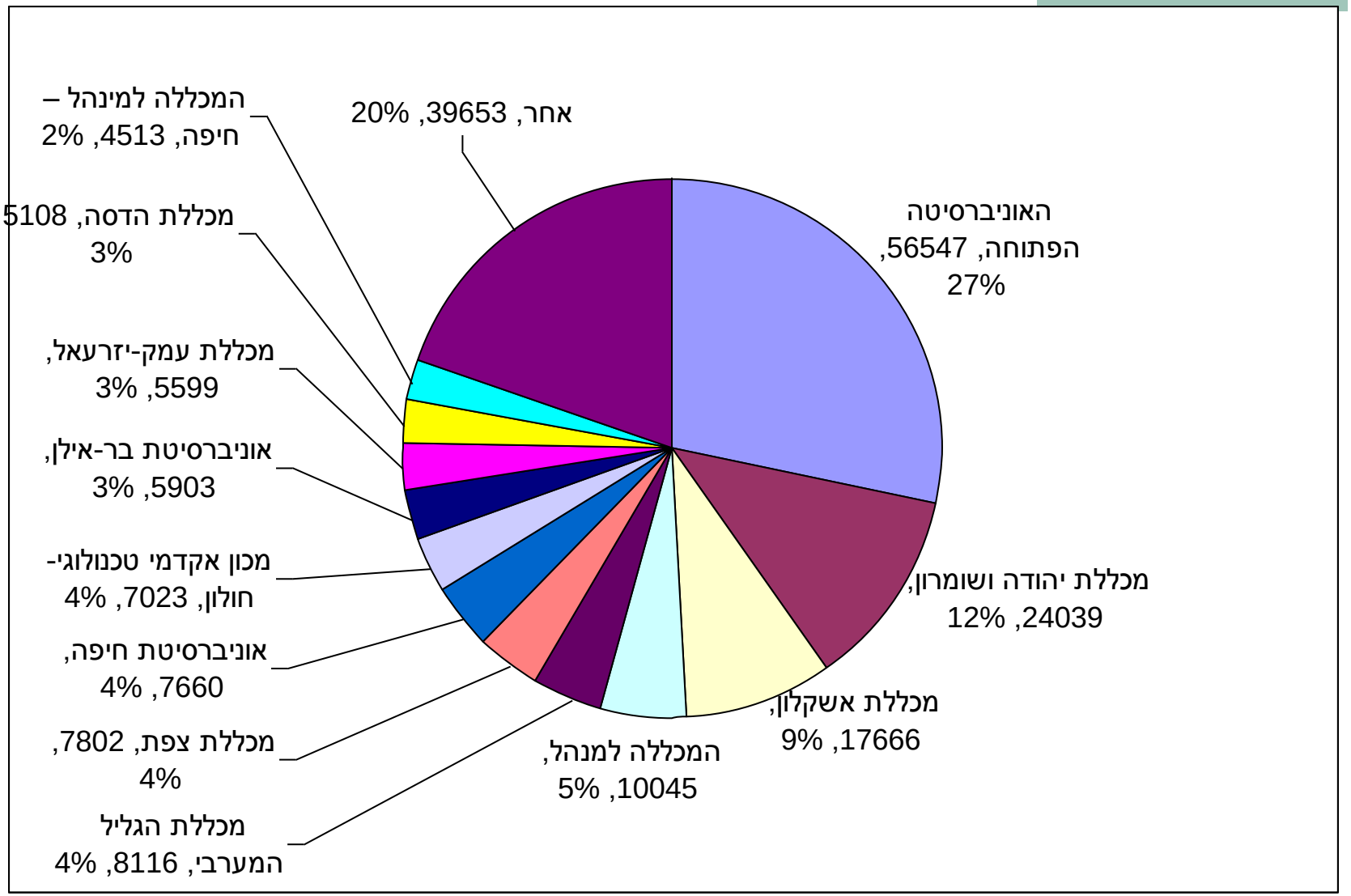
- **מבנה**
  - בבחינה שאלות **השלמת משפטים, ניסוח מחדש והבנת הנקרא**.
  - שאלות **ברֵרה** עם ארבע תשובות מוצעות.
  - אורך הבחינה הממוצע **כשעה**.

  - בחינה **ממוחשבת**.
  - בחינה **אדפטיבית**, לפי מודל IRT תלת פרמטרי*.
  - רמת היכולת מומרת לציון סופי בבחינה באמצעות **טבלת המרה***.
  - התאמת קושי השאלות ליכולת הנבחן מאפשרת לאמוד את רמת הידע שלו בעזרת **מספר שאלות קטן** (21,28 שאלות; ממוצע 23).

* ידובר בהמשך.

תוצאות

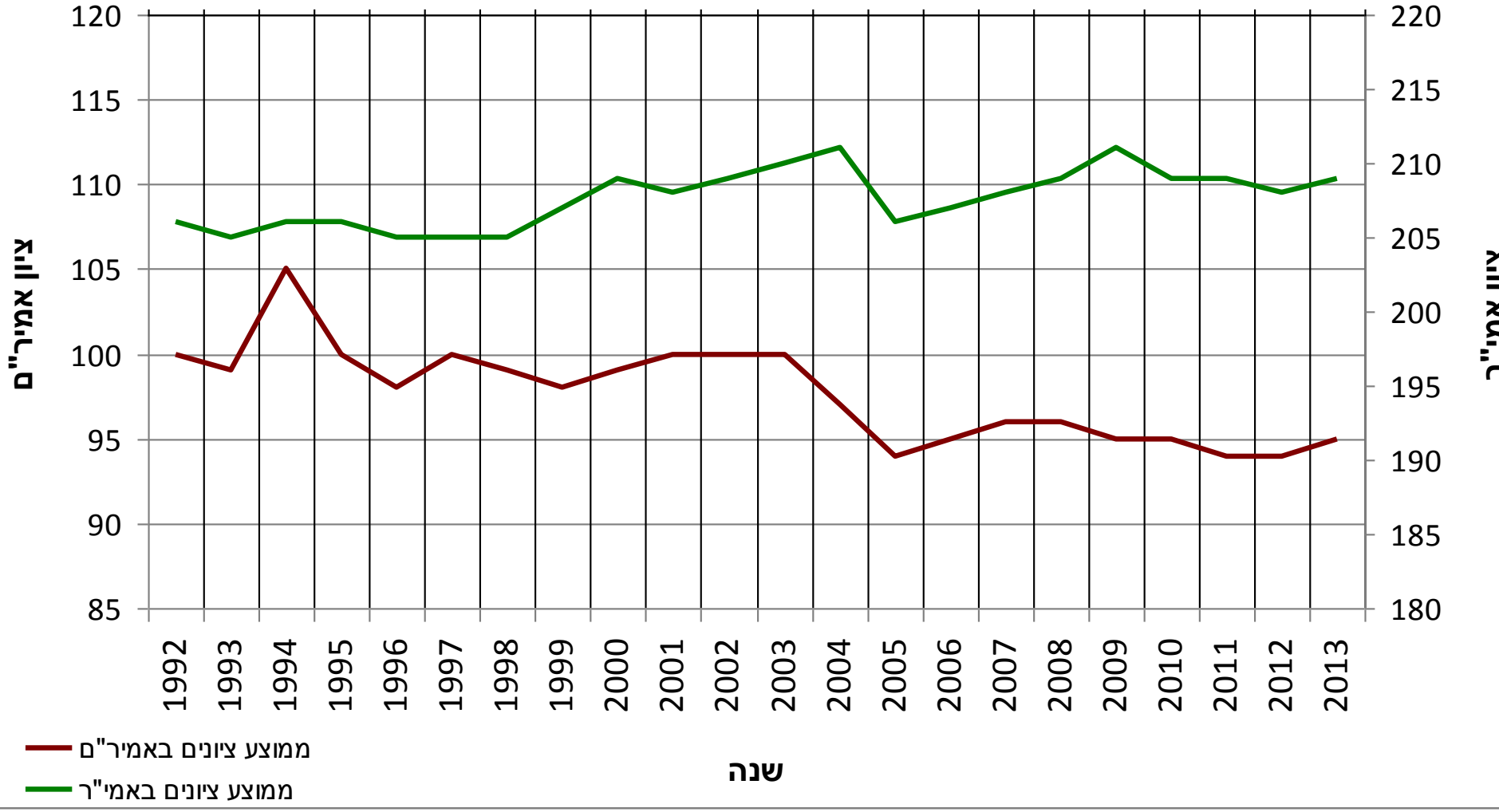
שכיחויות נבחנים על-פי מוסד אקדמי מעבר לשנים



שכיחויות נבחנים באמיר"ם מול אמי"ר לפי שנה

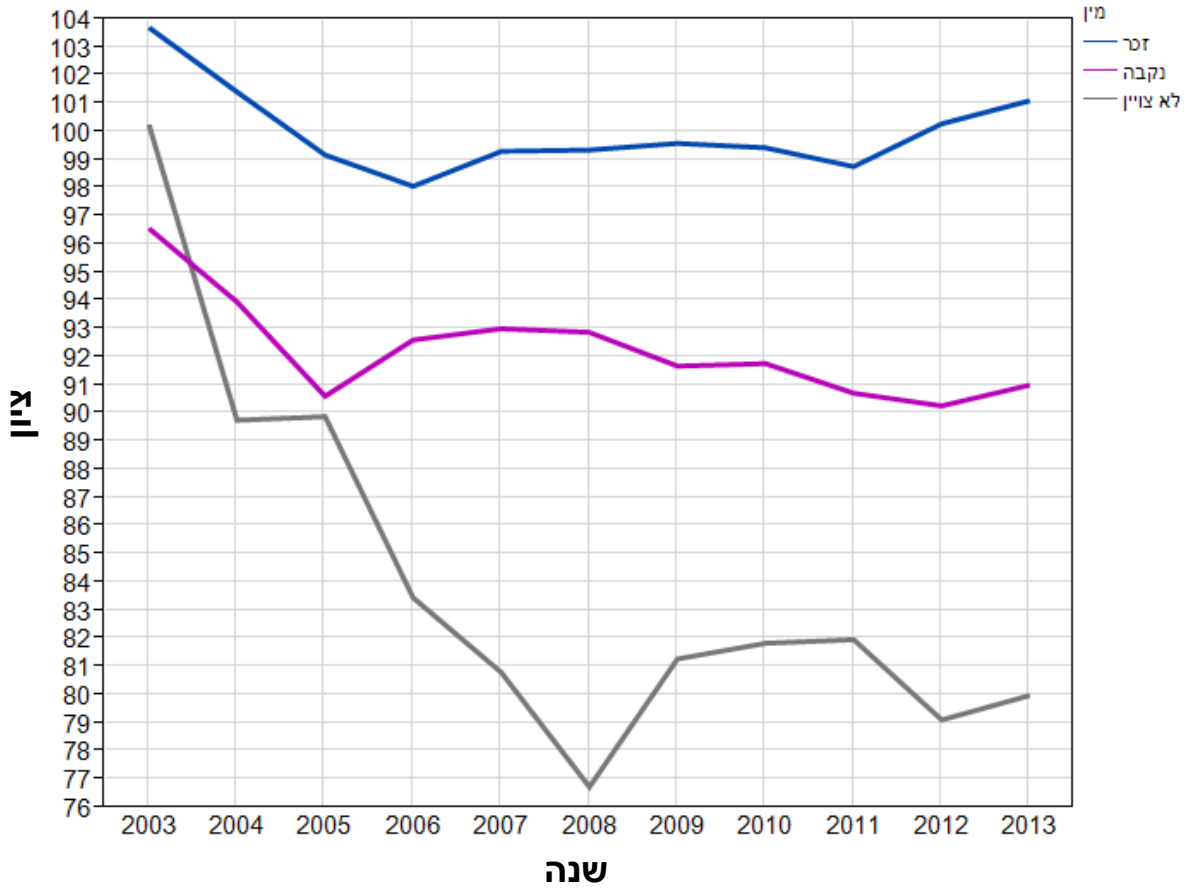
ציוני אמיר"ם לעומת אמי"ר לפי שנה

[הבדל של 100 בסולם]



מגדר, ציון ממוצע לפי שנה 2003-2013

(לפני שנת 2003 הנבחנים לא נתבקשו לציין מגדר)

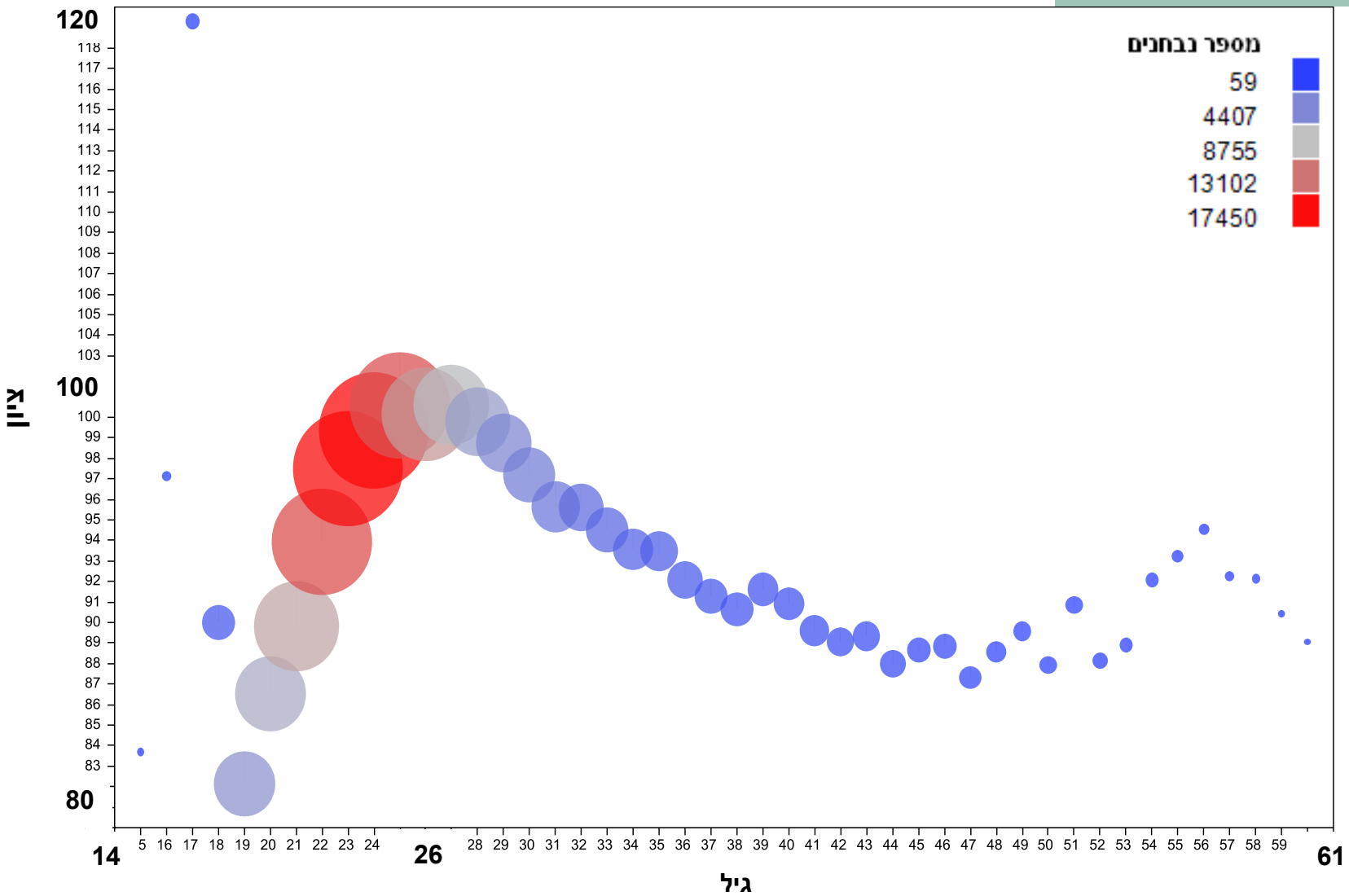


רוב הנבחנים הן נבחנות (נשים: 89,391, גברים: 66,779).

ציונן של הנשים (M=92.0, SD=24.3) נמוך במובהק מציון הגברים (M=99.9, SD=25.9) בכ-0.4 ס.ת.

גיל, שכיחות וציון ממוצע מעבר לשנים 2003-2013

(לפני שנת 2003 הנבחנים לא נתבקשו לציין גיל)



שכיחות, ממוצע וסטיית תקן של הציון לפי נוסח מקובץ מעבר לשנים

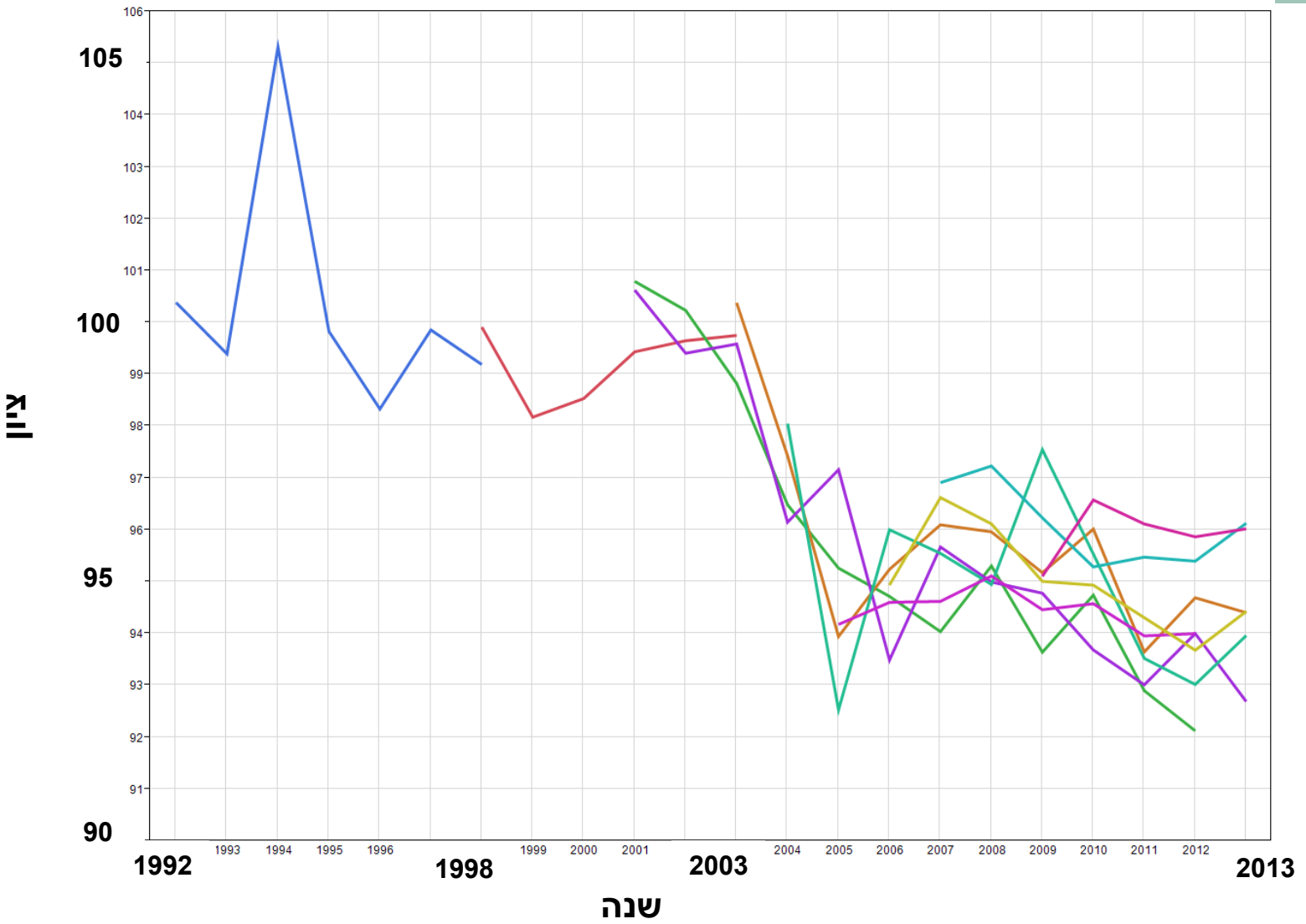
ניתן לראות כי ציון הבחינה הממוצע וסטיית התקן יציבים מעבר לנוסחים השונים, החל מהנוסח ה-3.

		ציון	
נוסח	N	ממוצע	ס.ת.
1	16,072	100.1	20.4
2	24,805	99.0	20.8
3	18,049	95.8	24.4
4	20,274	95.5	24.6
5	27,397	95.6	25.0
6	22,165	94.7	25.7
7	17,570	94.5	25.5
8	19,668	95.0	25.4
9	20,149	96.0	25.6
10	13,529	96.0	26.6

נוסח ישן

נוסח חדש

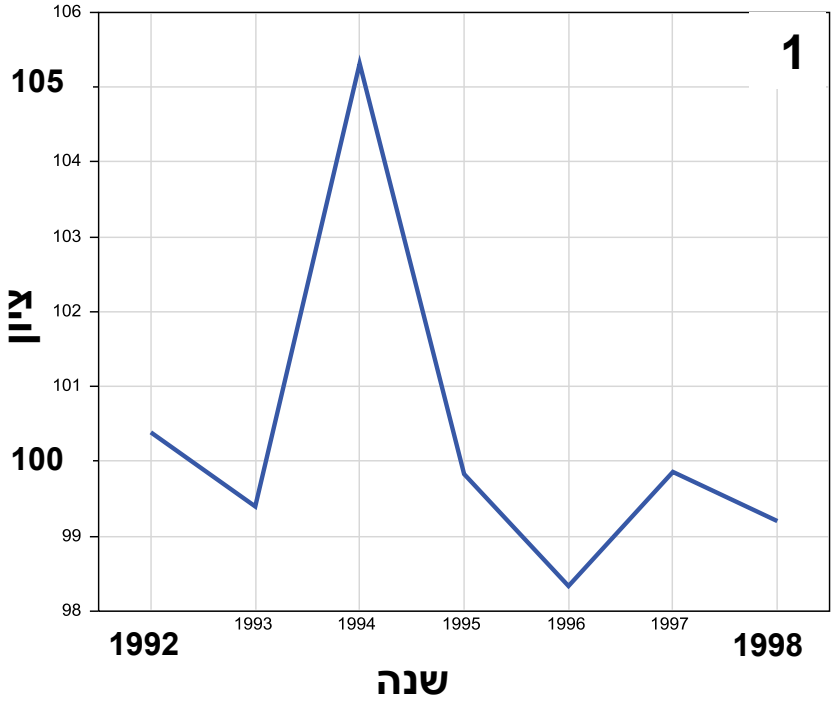
ציון ממוצע על-פני ציר הזמן
לפי נוסח מקובץ 1992-2013

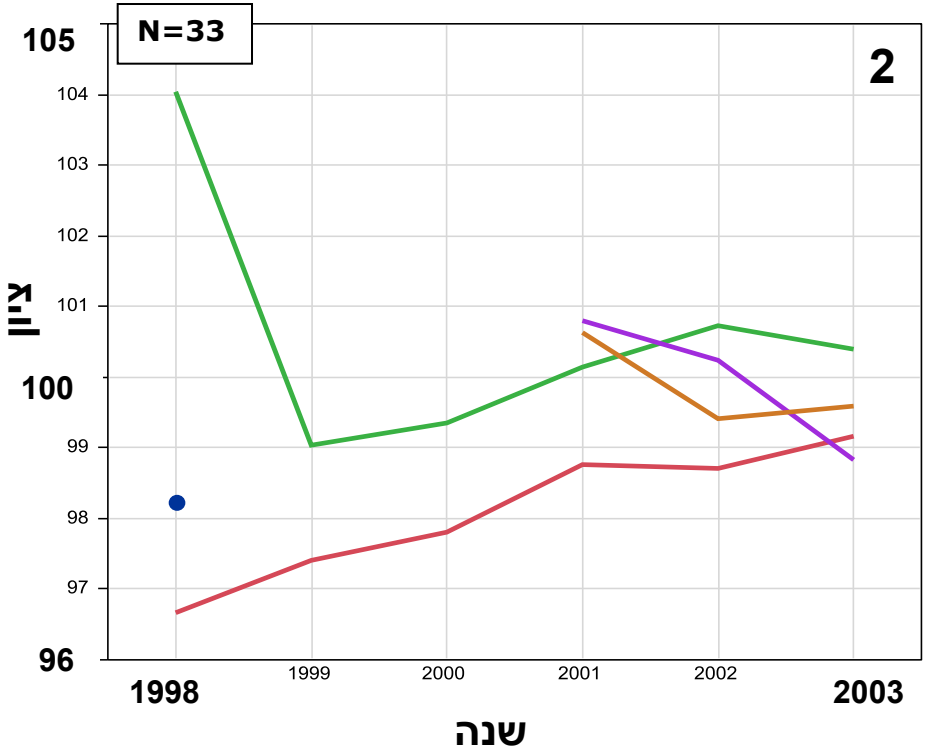


ציון ממוצע על-פני ציר הזמן לפי נוסח

העשור הראשון

בעשור הראשון ניכרה תנודה משמעותית בציון הממוצע משנה לשנה.

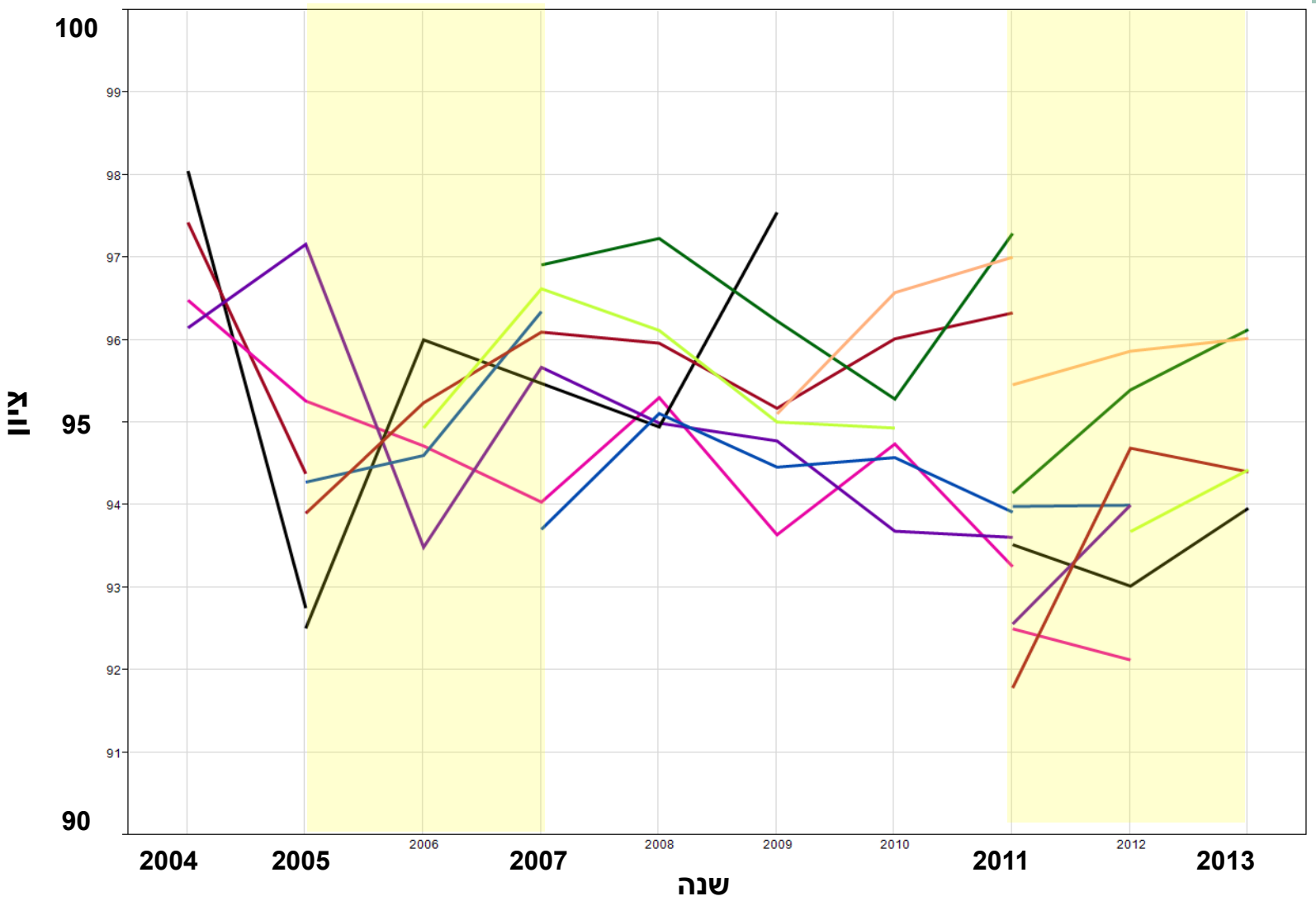




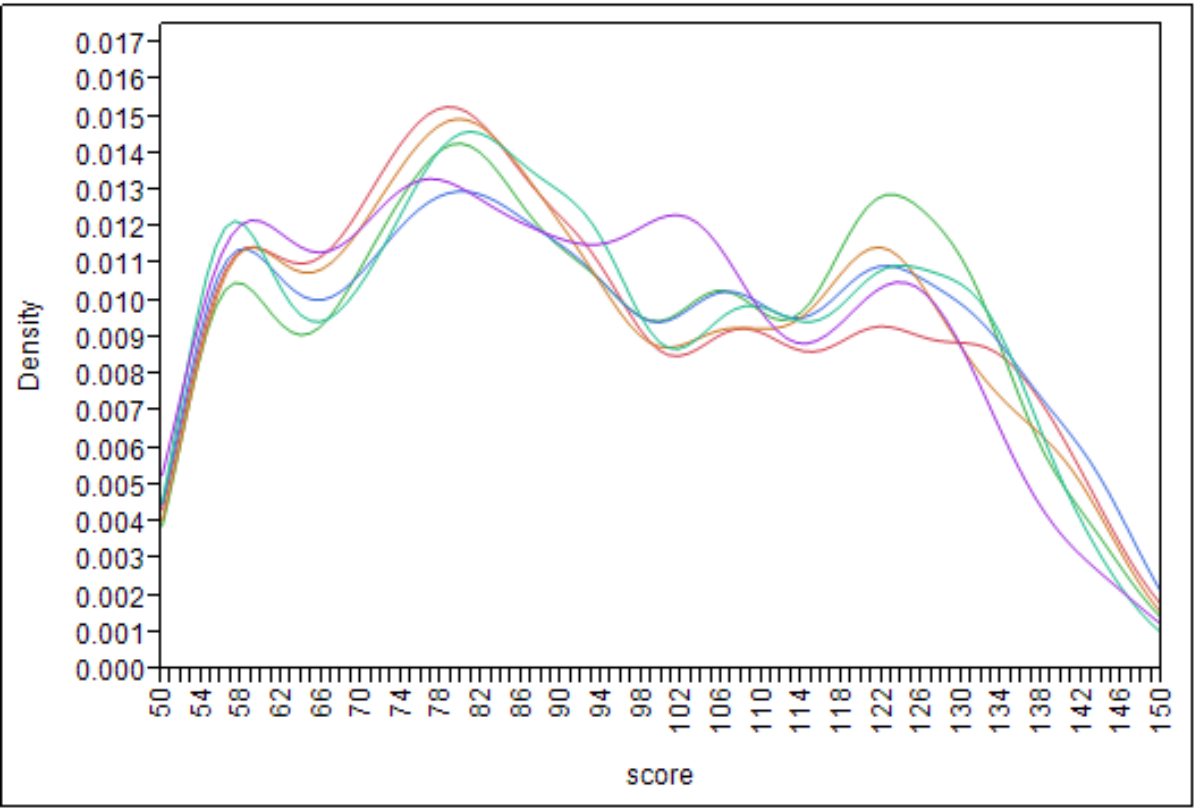
לקראת העשור השני, כל הנוסחים המוצגים כאן הורדו מהמחזור.

ציון ממוצע על-פני ציר הזמן לפי נוסח

העשור השני



שקילות נוסחים 2013



Level				Mean
א'	A			96.1
ב'	A	B		96.0
ג'		B	C	94.4
ד'		B	C	94.4
ה'			C	94.0
ו'			C	92.7

נמצא הבדל מובהק בין נוסחי הבחינה, אך האפקט זעום:
F (5, 21707) = 7.06, p < .001, *adj. r2* = .01.

טבלאות המרה: מתטא לציון הבחינה

- הפרמטר תטא (θ) מייצג את **יכולת** הנבחן, במקרה זה בתחום האנגלית.
- הפרמטר מחושב על סמך מודל IRT לוגיסטי תלת פרמטרי (3PL) בשיטת הנראות המרבית.
- הטבלאות ממירות

  רמת יכולת (ציון תטא, לרוב בטווח של 3- עד 3)

  לציון סופי בבחינה (על סולם של 50 עד 150)
- טבלת ההמרה נקבעה על סמך מחקרים אמפיריים (רפ, 1995, 1996) והחלטות לגבי ערכים ספציפיים: "**כללי עצירה**".

אחוזי התכנסות לפי נוסח 2012-2013

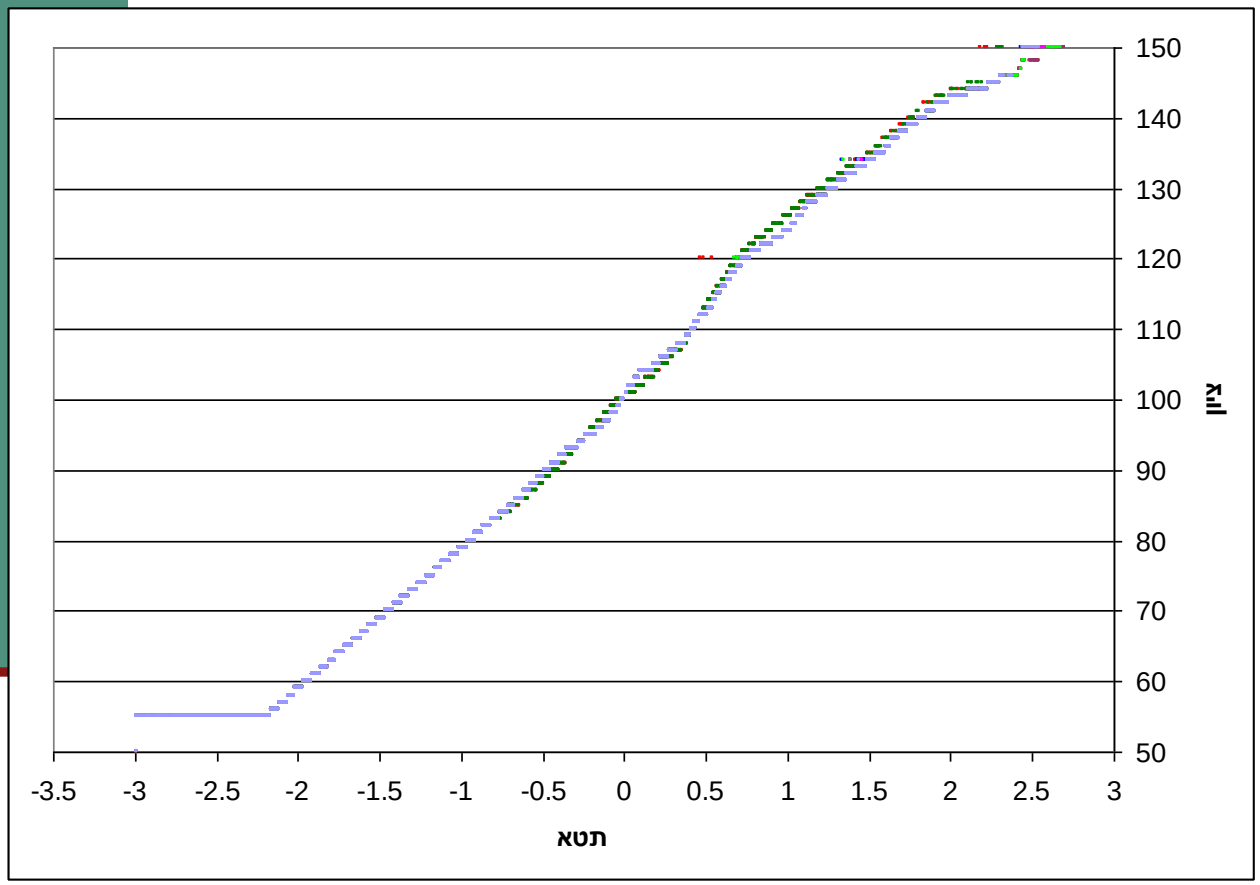
- P-var היא טעות המדידה, נאמדת לאחר כל פריט במבחן.

נוסח	אחוז הנבחנים שציונם התכנס
א'	81
ג'	80
ה'	80
ו'	80
ז'	80
ח'	80
ב'	79
ד'	77

התכנסות מוגדרת כ-P-var=<.08.

לעיתים, המבחן יסתיים ללא התכנסות בנקודת החתך (אך במרחק סביר ממנה).

טבלאות המרה
נוסחים שהועברו ב-2013



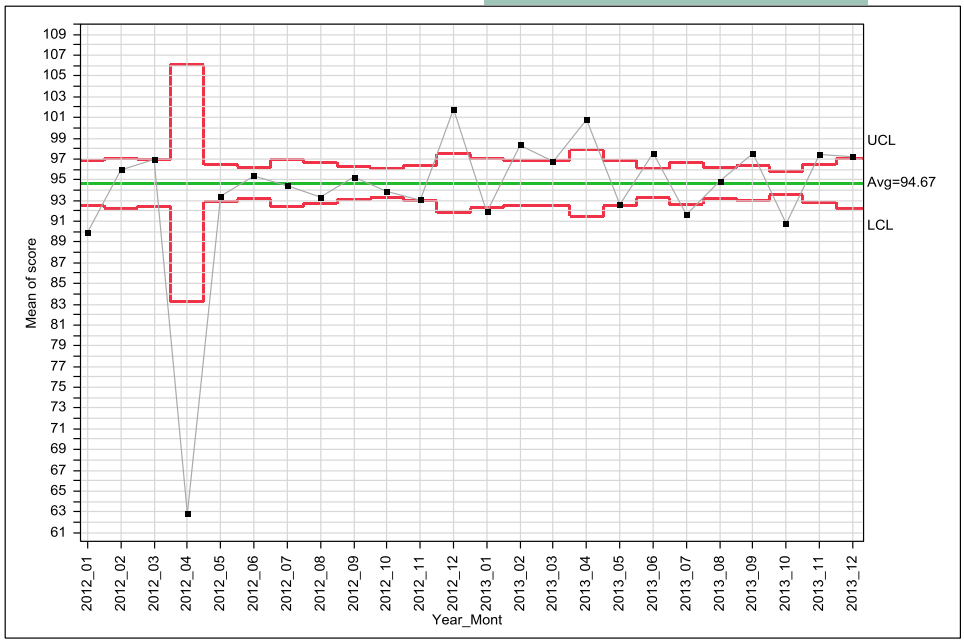
הקשר בין תטא לציון כמעט מושלם (r = .99).

בתרשים מוצגים כל הנוסחים שבשימוש כיום, כל נוסח בצבע קו שונה.

ניתן לראות את הפעלתם של "**כללי עצירה**".

תרשימי בקרה על הציונים הגולמיים בשנים 2012-2013

ממוצע



סטיית תקן

דין

מסקנות

סטטיסטיקה תיאורית

- ישנו שימוש גובר בבחינה במוסדות חדשים לצד הוותיקים (למשל, נבחני האו"פ).
- לצד העלייה העקבית במספר הנבחנים, ניכרת ירידה עקבית בציון הממוצע בבחינה.
- שכיחות הנבחנים באמיר"ם עולה כיום על שכיחות הנבחנים באמי"ר; הציון הממוצע באמיר"ם נמוך מזה של אמי"ר.
- גיל הנבחנים הממוצע הינו 26.6 (ס.ת. 7.6); וגברים מראים בממוצע הישגים גבוהים מנשים (בכ-0.4 ס.ת.).

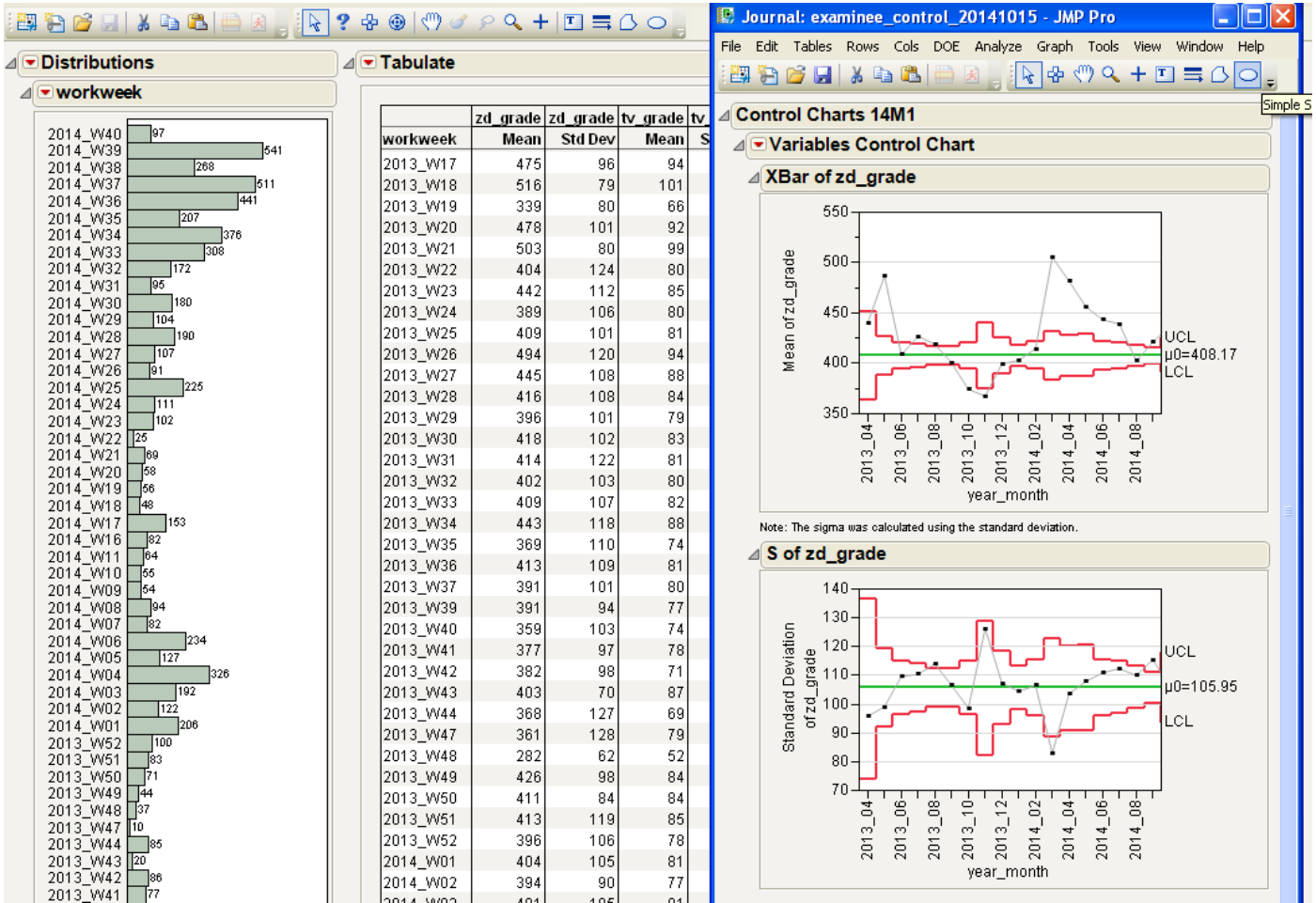
מסקנות (2)

- **בקרת איכות**
  - הציון הממוצע וסטיית התקן בנוסחים השונים די יציבים.
  - התכנסות הבחינה עומדת על כ-80 אחוז בממוצע.
  - תרשימי בקרת איכות מאפשרים מעקב ויזואלי נוח אחר ממוצעי הציונים.

המחקר הנוכחי מרחיב את היריעה בדרך לתפיסה כוללת של בקרת איכות של מבחנים המועברים תכופות לקבוצות נבחנים קטנות, תוך נתינת מענה גם **למבחנים אדפטיביים** בנוסף למבחנים לינאריים.

פלט המערכת הממוכנת לבקרה מתמשכת על בחינת המימ"ד

אחד מתוצרי המחקר המתוכננים הינו **מערכת ממוכנת אוטומטית** לבקרת הציונים טרם דיווחם, שתפעל באופן שבועי.



workweek	zd_grade Mean	zd_grade Std Dev	tv_grade Mean	tv_ S
2013_W17	475	96	94	
2013_W18	516	79	101	
2013_W19	339	80	66	
2013_W20	478	101	92	
2013_W21	503	80	99	
2013_W22	404	124	80	
2013_W23	442	112	85	
2013_W24	389	106	80	
2013_W25	409	101	81	
2013_W26	494	120	94	
2013_W27	445	108	88	
2013_W28	416	108	84	
2013_W29	396	101	79	
2013_W30	418	102	83	
2013_W31	414	122	81	
2013_W32	402	103	80	
2013_W33	409	107	82	
2013_W34	443	118	88	
2013_W35	369	110	74	
2013_W36	413	109	81	
2013_W37	391	101	80	
2013_W39	391	94	77	
2013_W40	359	103	74	
2013_W41	377	97	78	
2013_W42	382	98	71	
2013_W43	403	70	87	
2013_W44	368	127	69	
2013_W47	361	128	79	
2013_W48	282	62	52	
2013_W49	426	98	84	
2013_W50	411	84	84	
2013_W51	413	119	85	
2013_W52	396	106	78	
2014_W01	404	105	81	
2014_W02	394	90	77	