

# בקרת איכות דו-רמתית על ציונים לפי משתנים דמוגרפיים

דביר קלפר  
אבי אללוף  
אליוט טורוול  
כרמל אורן  
מרינה פרונטון

מאי 2015



מרכז ארצי לבחינות ולהערכה (ע"ר)

NATIONAL INSTITUTE FOR TESTING & EVALUATION

المركز القطري للامتحانات والتقييم

מיסודן של האוניברסיטאות בישראל

**דוח מרכז 415**  
**ISBN:978-965-502-190-5**

## תוכן עניינים

3.....	1	תמצית	1
4.....	2	מבוא	2
4.....	2.1	בקרת איכות	2.1
5.....	2.2	מטרות המחקר וחשיבותו	2.2
6.....	3	שיטה	3
6.....	3.1	כלי המדידה	3.1
6.....	3.2	האוכלוסייה	3.2
7.....	3.3	מודלים דו-רמתיים	3.3
7.....	3.3.1	ניתוח שונות עם אפקטים מקריים	3.3.1
8.....	3.3.2	רגרסיה של הממוצעים	3.3.2
8.....	3.3.3	רגרסיה עם מקדמים מקריים	3.3.3
10.....	4	תוצאות	4
10.....	4.1	סמטיסטיקה תיאורית	4.1
10.....	4.1.1	הקריטריון	4.1.1
10.....	4.1.2	הקריטריון בחלוקה לחזאים השונים	4.1.2
14.....	4.1.3	החזאים ברמה השנייה	4.1.3
15.....	4.2	מודלים דו-רמתיים	4.2
15.....	4.2.1	ניתוח שונות עם אפקטים מקריים	4.2.1
16.....	4.2.2	רגרסיה של הממוצעים	4.2.2
18.....	4.2.3	רגרסיה עם מקדמים מקריים	4.2.3
20.....	4.3	תיקוף	4.3
22.....	5	דיון	5
22.....	5.1	אחרית דבר	5.1
23.....	5.2	הצעות למחקרי המשך	5.2
24.....	6	מקורות	6

## רשימת לוחות

- לוח 1 – סטטיסטיקה תיאורית עבור הציון הכולל (הקריטריון) ..... 10
- לוח 2 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מין ..... 10
- לוח 3 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה רמת הכנסה ..... 11
- לוח 4 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אב ..... 12
- לוח 5 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אם ..... 13
- לוח 6 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מספר היבחנות ..... 14
- לוח 7 – סטטיסטיקה תיאורית של הציון הכללי עבור החזאים ברמה השנייה ..... 14
- לוח 8 – ניתוח שונות עם אפקטים מקריים ..... 15
- לוח 9 – תוצאות שש רגרסיות עם חזאי יחיד לניבוי ממוצע הציון הכולל ..... 16
- לוח 10 – תוצאות הרגרסיה של ממוצעי החזאים לניבוי ממוצע הציון הכולל – כל החזאים ..... 17
- לוח 11 – תוצאות הרגרסיה של ממוצעי החזאים לניבוי ממוצע הציון הכולל – המודל הסופי ... 17
- לוח 12 – תוצאות שש רגרסיות עם מקדמים מקריים שבכול אחת מהן חזאי בודד ..... 19
- לוח 13 – תוצאות רגרסיה עם מקדמים מקריים שכללה את כל החזאים ביחד ..... 19
- לוח 14 – מספר נוסח, ממוצע מנובא, ממוצע בפועל והפער ביניהם עבור 34 נוסחי התיקוף ..... 20

## 1. תמצית

חישוב ציונים בבחינות רחבות היקף ועתירות סיכון הינו תהליך רב-שלבי מורכב ולכן עשויות לחול בו טעויות שמחירן עלול להיות גבוה לכל הגורמים המעורבים בבחינה. מכיוון שכך, חשוב לבצע פעולות של בקרת איכות קפדנית על ציונים לפני דיווחם לנבחנים ולמוסדות (Allalouf, 2007; Kolen & Brennan, 2014). אחת משיטות הבקרה היא להיעזר בנתונים ידועים מהעבר על קבוצות נבחנים ולהשוות אותם לציוני הקבוצות שהתקבלו בחישוב החדש. במחקרים רבים נמצא קשר בין נתוני רקע דמוגרפיים (מין, גיל, השכלת הורים, הכנסה וכו') להישגים בבחינות שונות (למשל Liu et. al. 2012), כולל הבחינה הפסיכומטרית (סער ואורן 2014). קיומו של קשר זה מאפשר לנבא את ציון הבחינה, הן זה של נבחן בודד והן הממוצע של קבוצת נבחנים וזאת על מנת לאמת את הציונים שהתקבלו בתהליך החישוב. צפוי כי ניבוי פרמטרים קבוצתיים (של הנבחנים בנוסח מסוים במועד בחינה מסוים), כמו ממוצע וחציון, יהיה מדויק יותר (וטעות הניבוי תהיה קטנה יותר), מאשר ניבוי ציון של נבחן בודד. מטרת המחקר הנוכחי היא לבדוק ולהשוות את השיטות של שלושה מודלים רב-רמתיים לניבוי הציון בבחינה הפסיכומטרית על בסיס משתני רקע שונים לצורך בקרת איכות של תהליך חישוב הציונים.

בהתאם לכך בוצע ניתוח היררכי לינארי בשתי רמות כדי לחקור את הקשר של נתוני הרקע לציון הכללי מעבר לנוסחים, כאשר הרמה הראשונה היא הנבחן הבודד, והרמה השנייה היא נוסח הבחינה.

שלושת המודלים שנבדקו הם:

1. ניתוח שונות עם אפקטים מקריים
2. רגרסיה של הממוצעים
3. מודל רגרסיה עם מקדמים מקריים

מודל 1 שימש כבסיס להשוואה עם שני המודלים האחרים. נמצא כי ברמת הנוסח יכולת ניבוי הממוצע היא טובה מאוד (מודל 2 הצליח להסביר 69% מהשונות של ממוצעי הנוסחים). עוד נמצא כי היכולת לנבא ציון של נבחן בודד היא חלשה יחסית (מודל 3 הצליח להסביר רק 20% מהשונות של הציון של נבחן בודד). מודל 2 עם יכולת הניבוי הטובה יותר נבדק על בסיס נתונים שכלל נוסחים שונים מאלה שנכללו במדגם חישוב הפרמטרים ונמצא תקף.

מחקר זה הינו מחקר ראשון המשתמש בניתוח רב-רמתי כדי לחקור את הקשר בין משתני הרקע של נבחנים לבין הציון הכללי בבחינה הפסיכומטרית. תוצאות המחקר מעודדות ונותנות ביד החוקר כלי רב-עוצמה לבקרת איכות של ציונים בבחינה זו כמו גם במגוון רחב של מבחנים שבהם יש בידי החוקר נתונים דמוגרפיים של הנבחנים. בקרת האיכות יכולה להתבצע הן לאחר דיווח הציונים במטרה לבקר את איכות תהליך חישוב הציונים והכיוול, והן לפני דיווח הציונים כדי לאתר בעיות בזמן אמת.

## 2.1. בקרת איכות

בקרת איכות מפותחת וחשובה במיוחד בדיסציפלינות שבהן לטעויות מחיר כבד, ולכן בקרת איכות בהן 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 החומרה ותוכנה. בקרת איכות על מוצרים תעשייתיים החלה לפני כמאה שנה (Shewhart, 1931). בקרת איכות בתעשייה מתבצעת לרוב על דגימות שיטתיות הנלקחות תוך כדי תהליך הייצור. ריכוז שיטות סטטיסטיות לבקרת איכות מופיע אצל Montgomery (2009).

בקרת האיכות הנערכת על מבחנים שונה באופייה מבקרת איכות על מוצר תעשייתי. אחד ההבדלים קשור בכך שמוצר תעשייתי הוא מוחשי ולפחות חלק מהפגמים, אם יש כאלה, נראים לעין המתבונן ויתגלו לכן בטווח הקצר, לעיתים תוך כדי השימוש במוצר. כשמדובר במבחנים חינוכיים ופסיכולוגיים, התוצרים הינם מסוג אחר, תוצרי תהליכי מדידה האמורים לייצג תכונות, יכולות וכישורים של הפרט שאינם מוחשיים. אלה הינם לרוב תוצרים מספריים: ציונים ומדדים סטטיסטיים. אם נפלה בהם שגיאה, מירב הסיכויים שהיא תתגלה מאוחר מאוד או לא תתגלה לעולם. בנוסף, בניגוד לבקרת איכות על מוצרים תעשייתיים שנערכת על מדגם מתוך המוצרים בתכיפות גבוהה, בקרת איכות על מבחנים נערכת לרוב על האוכלוסייה כולה, כל תקופת זמן קצובה.

חישוב ציונים בבחינות סטנדרטיות רחבות היקף הינו תהליך מורכב ורב-שלבי המסתמך על הנחות סטטיסטיות מורכבות ונתון לטעויות. מחיר הטעויות עלול להיות גבוה לכל המעורבים בתהליך, בפרט כשמדובר בבחינות עתירות סיכון (high stakes tests). מכיוון שכך, יש לבצע בקרות איכות קפדניות על ציונים לפני דיווחם (Allalouf, 2007; Kolen & Brennan, 2014) ולעשות זאת ביעילות ובמהירות, שכן חלון הזמן שבין חישוב בציונים ודיווחם הוא מוגבל. אחת האפשרויות לבצע בקרה היא להשוות את ממוצע הציונים שהתקבלו עבור קבוצת נבחנים לממוצע הציונים שהיינו מנבאים לה בעזרת נתוני רקע שונים. הניבוי של ממוצע ציוני קבוצה יכול להיעשות בכמה דרכים, למשל, על סמך ביצוע קבוצה דומה במועד בחינה מקביל בשנה הקודמת, בהנחת יציבות מועדי בחינה מקבילים.

מחקרים רבים מצאו שקיים קשר בין נתונים דמוגרפיים (מין, גיל, השכלת הורים ועוד) להישגים לימודיים וציוני בחינות יכולת (למשל Liu et. al. 2012). בדוח של ה-College Board (2013) המציג את הישגי הנבחנים ב-SAT לפי פילוח של משתנים דמוגרפיים רבים, ביניהם מין, מוצא ושפת אם, ניתן לראות את ההבדלים בציונים בין הקבוצות השונות. נתונים דומים קיימים גם לגבי הבחינה הפסיכומטרית (סער ואורן 2014). ידיעת הקשרים בין המאפיינים הדמוגרפיים לבין

הציון מאפשרת לנבא את הציונים, הן של נבחן בודד, והן של קבוצת נבחנים. נצפה כי ניבוי פרמטרים קבוצתיים (למשל של קבוצת נבחנים בנוסח מסוים במועד מסוים) יהיה מדויק יותר, מאשר ניבוי ציון של נבחן בודד. כן צפוי שניבוי ערכים מרכזיים כמו ממוצע וחציון יהיה מדויק יותר מניבוי מדדי פיזור וצורת התפלגות.

בתהליך בקרת האיכות משווים את הציונים המחושבים לציונים המנובאים בעזרת משתני הרקע, וכאשר מוצאים פער גדול ביניהם נהוג לחזור ולבדוק את תהליך חישוב הציונים. לכן, כלי המאפשר לאמוד "ציונים מנובאים" באופן מדויק עשוי לסייע לבקרה ולהדליק "נורת אזהרה" במקרה של פער גדול בין ציונים מחושבים לציונים מנובאים. אנשי המקצוע אמורים להסתייע באזהרה זו, לצד אזהרות אחרות, אם יש, כדי לאתר בעיות בזמן אמת ואת הסיבות להן. ניתן לומר כי בקרת איכות מעלה את רמתו המקצועית של המבחן ויש לעקוב אחריה ולשפרה.

## 2.2. מטרות המחקר וחשיבותו

מטרת המחקר הנוכחי היא לבדוק את הישימות של שלושה מודלים רב-רמתיים לניבוי הציון בבחינה הפסיכומטרית על בסיס משתני רקע שונים, ואת האפשרות להסתמך על המודל הטוב מביניהם לביצוע בקרת איכות של תהליך חישוב הציונים (Kidwell, Mossholder & Bennet, 1997; Gamoran et. al., 1997; Rumberger, 1995). זהו מחקר ראשון שנערך במרכז הארצי לבחינות ולהערכה (מאל"ו) המשתמש בניתוח רב-רמתי לחקירת הקשר בין משתנים דמוגרפיים של נבחנים לבין הציון הכללי בבחינה הפסיכומטרית. תוצאות המחקר עשויות לתת ביד החוקר כלי רב-עוצמה לבקרת איכות של ציונים בבחינות. בקרת האיכות יכולה להתבצע בדיעבד, על מבחנים שכבר הועברו, כדי לבקר את איכות חישוב הציונים והכיול, ובאופן שוטף, לאחר חישוב ציוני הבחינה וטרם דיווחם לנבחנים ולמוסדות המשתמשים, כדי לאתר בעיות בזמן אמת.

### 3.1. כלי המדידה

#### קריטריון לניבוי:

הבחינה הפסיכומטרית היא כלי לניבוי סיכויי ההצלחה בלימודים במוסדות להשכלה גבוהה, והיא משמשת את מוסדות הלימוד במיין מועמדיהם לחוגים השונים. הבחינה מועברת בחמישה מועדים שונים בשנה, כאשר בכל מועד יש מספר נוסחים בשפות שונות. בבחינה ניתן ציון בכל אחד מחלקיה: כמותי, מילולי ואנגלית וציון כללי. מחקר זה מתמקד בניבוי הציון הכללי בלבד (שבו לציון המילולי והכמותי משקל כפול מזה של האנגלית) בקרב הנבחנים בשפה העברית.

#### חזאים (משתני רקע דמוגרפיים):

להלן שישה משתנים. כולם, למעט האחרון, מדווחים על ידי הנבחן, בעת הרשמתו לבחינה.

- גיל – משתנה רציף
- מין (זכר=0, נקבה=1) – משתנה דיכוטומי
- רמת הכנסה של המשפחה (1=גבוהה בהרבה מהמוצע...6=נמוכה בהרבה מהמוצע)
- השכלת אב (1=לא למד...7=תואר שני ומעלה)
- השכלת אם (1=לא למד...7=תואר שני ומעלה)
- מספר היבחנות – 1 ומעלה

אנו מתייחסים אל ארבעת המשתנים האחרונים כאל משתני רווח.

### 3.2. האוכלוסייה

המדגם מכיל את כל הנבחנים בבחינה הפסיכומטרית אשר:

- נבחנו בין שנת 2000 לשנת 2013 (כולל)
  - נבחנו בשפה העברית
  - נבחנו בנוסח שבו כ-1,000 איש לפחות
  - לא הכילו חסרים באף אחד ממשתני הרקע (ראה לעיל)
- בסה"כ התקבל מדגם של 421,515 איש שניגשו ל-126 נוסחים (ממוצע של 3,345 לנוסח). נעיר שמספר הנבחנים המקיימים את שלושת התנאים הראשונים הוא 649,797, כך שהתנאי האחרון מצמצם אוכלוסייה זו לכ-66%.



מדגם זה פוצל לשניים : המדגם ששימש לבניית המודל והמדגם ששימש לתיקופו

- במדגם ששימש לבניית המודל נכללו 340,669 איש שנבחנו בשנים 2000-2010 ב-92 נוסחים
- במדגם התיקוף נבחנו 80,846 איש בשנים 2011-2013 ב-34 נוסחים

### 3.3 מודלים דו-רמתיים

בוצע ניתוח היררכי לינארי (Hierarchical Linear Modeling – HLM) בשתי רמות כדי לחקור את הקשר של נתוני הרקע לציון הכללי מעבר לנוסחים כאשר :

- רמה 1 : הנבחן הבודד
  - רמה 2 : נוסח בחינה
- כל אחד מהחזאים הוכנס הן כמשתנה רמה ראשונה והן כמשתנה רמה שנייה, כאשר אנו משתמשים בממוצע המשתנה בנוסח כמשתנה רמה שנייה.

נציין שלצורך ביצוע המודלים השתמשנו בתוכנת SAS ובפרוצדורה Proc Mixed (בספרו Wang, 2011 מסביר את השיטה ב-SAS).

בהתאם למקובל במחקרים רב-רמתיים לצורך בקרת איכות (Wei, 2013) נבדקו שלושה מודלים : ניתוח שונות עם אפקטים מקריים, רגרסיה של הממוצעים ומודל רגרסיה עם מקדמים מקריים.

#### 3.3.1 ניתוח שונות עם אפקטים מקריים

להלן משוואות המודל.

- רמה ראשונה :  $Y_{ij} = \beta_{0j} + r_{ij}$
  - רמה שנייה :  $\beta_{0j} = \gamma_{00} + u_{0j}$
- כאשר :

- $Y_{ij}$  הוא הציון של נבחן  $i$  בנוסח  $j$
- $\beta_{0j}$  הוא הציון הממוצע של נבחנים בנוסח  $j$
- $r_{ij}$  הוא השארית הקשורה בנבחן  $i$  בנוסח  $j$  ומניחים שהיא מתפלגת נורמלית עם שונות  $\sigma^2$
- $\gamma_{00}$  הוא הממוצע הכולל (ממוצע של ממוצעי הנוסחים)

○  $u_{0j}$  הוא אפקט מקרי הקשור בנוסח  $j$  ומניחים שהוא מתפלג נורמלית עם שונות  $\tau_{00}$

כפי שניתן לראות ממשוואות המודל, מודל זה אינו כולל משתנים מנבאים בכלל. בהתאם לכך, מודל זה ישמש כבסיס להשוואה עם שני המודלים האחרים.

### 3.3.2. רגרסיה של הממוצעים

להלן משוואות המודל.

- רמה ראשונה:  $Y_{ij} = \beta_{0j} + r_{ij}$
- רמה שנייה:  $\beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j}$

כאשר:

- $Y_{ij}$  הוא הציון של נבחן  $i$  בנוסח  $j$
- $\beta_{0j}$  הוא הציון הממוצע של נבחנים בנוסח  $j$
- $r_{ij}$  הוא השארית הקשורה בנבחן  $i$  בנוסח  $j$  ומניחים שהיא מתפלגת נורמלית עם שונות  $\sigma^2$
- $G_j$  הוא חזאי משתנה רקע מרמה 2 לציון הממוצע בנוסח  $j$
- $\gamma_{00}$  הוא הממוצע הכולל מותנה בחזאי  $G_j$
- $\gamma_{01}$  הוא מקדם השיפוע ברגרסיה של החזאי  $G_j$  על  $\beta_{0j}$
- $u_{0j}$  הוא אפקט מקרי הקשור בנוסח  $j$  מותנה בחזאי  $G_j$

מודל זה כולל משתנים מנבאים ברמה השנייה בלבד ולכן המשוואה השנייה היא מודל לממוצע הנוסח.

### 3.3.3. רגרסיה עם מקדמים מקריים

להלן משוואות המודל.

- רמה ראשונה:  $Y_{ij} = \beta_{0j} + \beta_{1j}B_{1ij} + r_{ij}$
- רמה שנייה:  $\beta_{0j} = \gamma_{00} + u_{0j}$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

כאשר:

- $Y_{ij}$  הוא הציון של נבחן  $i$  בנוסח  $j$
- $B_{1ij}$  הוא חזאי משתני רקע מרמה 1
- $\beta_{1j}$  ו  $\beta_{0j}$  הם החותך והשיפוע ברגרסיה של החזאי מרמה 1 על  $Y_{ij}$
- $r_{ij}$  הוא השארית של החזאי מרמה 1
- $\gamma_{10}$  ו  $\gamma_{00}$  הם הממוצע הכולל והשיפוע הממוצע מעבר לנוסחים
- $u_{1j}$  ו  $u_{0j}$  הם האפקטים המקריים של החותך והשיפוע הקשורים בנוסח  $j$  ומניחים שהם בעלי מטריצת שונות משותפת הבאה:
 
$$\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}$$
- $\tau_{00}$  הוא השונות ברמה 1 של החותך,  $\tau_{11}$  הוא השונות ברמה 1 של השיפוע ו  $\tau_{01}$  ו  $\tau_{10}$  הם השונות המשותפת בין החותך והשיפוע ברמה 1

מודל זה מכיל משתנים מנבאים ברמה הראשונה בלבד ולכן הוא מציע ניבוי עבור הנבחן הבודד.

#### 4. תוצאות

##### 4.1. סטטיסטיקה תיאורית

##### 4.1.1. הקריטריון

לוח 1 מציג את השכיחות, הממוצע, סטיית התקן, המינימום והמקסימום של הציון הכולל (הקריטריון) בשתי הרמות.

לוח 1 – סטטיסטיקה תיאורית עבור הציון הכולל (הקריטריון)

מקסימום	מינימום	סטיית תקן	ממוצע	N	משתנה	רמה
800	227	100.5	565.9	340,669	הציון הכולל	רמה 1: רמת הנבחן
586.4	532.4	10.3	565.2	92	ממוצע	רמה 2: רמת הנוסח
108.3	93.3	3.1	100.2	92	סטיית תקן	
6,655	1,132	1,423.8	3,702.9	92	גודל המדגם	

##### 4.1.2. הקריטריון בחלוקה לחזאים השונים

בלוחות הבאים נציג סטטיסטיקה תיאורית עבור הציון הכולל (הקריטריון) לכול אחד מחמשת החזאים (המשתנה גיל הינו רציף ועבורו אנחנו לא מציגים סטטיסטיקה תיאורית).

לוח 2 מציג סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מין.

לוח 2 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מין

מקסימום ציון כולל	מינימום ציון כולל	סטיית תקן	ממוצע ציון כולל	אחוזים	N	קטגוריה	רמה
800	227	96.4	587.6	45	152,163	זכר	רמה 1: רמת הנבחן
800	231	100.4	548.4	55	188,506	נקבה	
608.7	556.8	10.6	587.2		92	זכר	רמה 2: רמת הנוסח
571.2	511.3	11.9	547.1		92	נקבה	

ניתן לראות בלוח 2 שיש יותר נשים מאשר גברים, כאשר הממוצע של הגברים גבוה מזה של הנשים. כפי שנראה בהמשך, לאחר הפעלת המודלים הסטטיסטיים, נמצא שבנוסחים בהם יש יותר נשים הממוצע של הנוסח גבוה יותר.

לוח 3 מציג סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה רמת הכנסה.

### לוח 3 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה רמת הכנסה

מקסימום ציון כולל	מינימום ציון כולל	סטיית תקן	ממוצע ציון כולל	אחוזים	N	קטגוריה	רמה
800	244	97.4	608.4	2.5	8,680	1=גבוה בהרבה	רמה 1: רמת הנבחן
800	227	92.6	606.1	19.8	67,601	2=גבוה	
800	231	95.5	572.9	42.3	144,191	3=גבוה במעט	
800	232	97.5	538.5	22.0	74,922	4=נמוך במעט	
793	279	100.6	525.1	9.9	33,881	5=נמוך	
800	251	105.8	508.2	3.3	11,394	6=נמוך בהרבה	
645.4	523.5	18.8	613.4		92	1=גבוה בהרבה	רמה 2: רמת הנוסח
625.1	562.3	10.8	604.9		92	2=גבוה	
592.0	545.1	10.7	572.0		92	3=גבוה במעט	
558.6	513.5	9.2	537.9		92	4=נמוך במעט	
555.8	499.3	10.1	524.5		92	5=נמוך	
540.1	471.3	15.2	508.7		92	6=נמוך בהרבה	

נשים לב שהסקלה (כפי שהיא מופיעה בשאלון המילוי העצמי) הפוכה, כך שקטגוריה 1 מציינת הכנסה גבוהה בהרבה מהממוצע בעוד שקטגוריה 6 מציינת הכנסה נמוכה בהרבה מהממוצע.

ניתן לראות בלוח 3 שקיים מתאם חיובי בין רמת ההכנסה לציון בבחינה.

לוח 4 מציג סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אב.

לוח 4 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אב

מקסימום ציון כולל	מינימום ציון כולל	סטיית תקן	ממוצע ציון כולל	אחוזים	N	קטגוריה	רמה
783	251	101.1	448.3	1.6	5,497	1=לא למד	רמה 1: רמת הנבחן
781	239	99.8	513.5	3.8	12,794	2=יסודית	
797	279	96.7	537.5	12.6	43,036	3=תיכונית חלקית	
800	227	96.8	543.9	25.8	88,000	4=תיכונית מלאה	
800	231	95.4	566.8	17.5	59,501	5=על תיכונית	
800	232	95.0	593.1	19.1	65,076	6=תואר ראשון	
800	232	91.7	605.8	19.6	66,765	7=תואר שני ומעלה	
525.8	404.2	25.0	453.2		92	1=לא למד	רמה 2: רמת הנוסח
543.9	481.5	14.4	513.1		92	2=יסודית	
568.9	510.6	10.9	537.8		92	3=תיכונית חלקית	
568.7	512.7	10.7	543.0		92	4=תיכונית מלאה	
588.9	532.6	10.1	566.0		92	5=על תיכונית	
609.5	563.7	10.3	591.9		92	6=תואר ראשון	
627.4	579.7	10.4	605.4		92	7=תואר שני ומעלה	

ניתן לראות בלוח 4 שקיים מתאם חיובי בין השכלת האב לציון בבחינה.

לוח 5 מציג סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אם.

לוח 5 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה השכלת אם

מקסימום ציון כולל	מינימום ציון כולל	סטיית תקן	ממוצע ציון כולל	אחוזים	N	קטגוריה	רמה
786	251	110.7	463.8	2.0	6,718	1=לא למד	
783	239	99.5	501.7	2.7	9,035	2=יסודית	
797	273	96.6	530.5	9.5	32,275	3=תיכונית חלקית	
795	232	96.5	545.4	28.2	95,954	4=תיכונית מלאה	רמה 1: רמת הנבחן
800	252	96.2	567.8	17.9	60,860	5=על תיכונית	
800	231	94.8	590.4	21.7	74,007	6=תואר ראשון	
800	227	91.6	605.5	18.1	61,820	7=תואר שני ומעלה	
570.2	391.4	27.5	447.7		92	1=לא למד	
535.9	470.7	14.7	500.3		92	2=יסודית	
552.6	497.2	11.0	530.2		92	3=תיכונית חלקית	
562.6	514.9	10.4	544.5		92	4=תיכונית מלאה	רמה 2: רמת הנוסח
588.2	532.4	11.2	566.2		92	5=על תיכונית	
609.4	552.7	10.3	589.4		92	6=תואר ראשון	
629.8	579.8	10.1	605.2		92	7=תואר שני ומעלה	

ניתן לראות בלוח 5 שקיים מתאם חיובי בין השכלת האם לציון בבחינה. בנוסף, לוח זה דומה מאוד ללוח 4 (השכלת אב), ואכן נראה בהמשך כי שני המשתנים הנ"ל מתחרים ביניהם והמודל בוחר רק באחד מהם.

לוח 6 מציג סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מספר היבחנות. נעיר שבלוח להלן אוחדו הקטגוריות של היבחנות 4 ויותר לכדי קטגוריה אחת וזאת בשל המספרים ההולכים וקטנים. איחוד זה נעשה אך ורק לצורך הצגה זאת ובמודל בהמשך יוכנס מספר היבחנות האמיתי.

**לוח 6 – סטטיסטיקה תיאורית של הציון הכללי עבור המשתנה מספר היבחנות**

מקסימום ציון כולל	מינימום ציון כולל	סטיית תקן	ממוצע ציון כולל	אחוזים	N	קטגוריה	רמה
800	231	101.4	554.6	68.9	234,891	היבחנות 1	רמה 1: רמת הנבחן
800	227	93.0	588.0	24.2	82,548	היבחנות 2	
800	286	93.4	597.9	5.2	17,789	היבחנות 3	
800	284	105.0	612.8	1.6	5,441	היבחנות +4	
575.8	518.0	10.1	554.0		92	היבחנות 1	רמה 2: רמת הנוסח
610.2	553.4	11.3	587.5		92	היבחנות 2	
627.6	549.5	15.5	597.7		92	היבחנות 3	
688.1	527.9	32.9	606.4		92	היבחנות +4	

כפי שניתן לראות בלוח 6 קיים מתאם חיובי בין מספר ההיבחנות לציון בבחינה.

**4.1.3. החזאים ברמה השנייה**

כאמור, אנו משתמשים בממוצע כל משתנה ברמה הראשונה כתזאי עבור הרמה השנייה. לוח 7 מציג את הסטטיסטיקה התיאורית של משתני הרקע עבור הרמה השנייה.

**לוח 7 – סטטיסטיקה תיאורית של הציון הכללי עבור החזאים ברמה השנייה**

מקסימום הממוצעים	מינימום הממוצעים	סטיית תקן של הממוצעים	ממוצע הממוצעים	משתנה
0.61	0.46	0.03	0.55	מין
23.05	19.03	0.98	21.52	גיל
3.49	2.85	0.09	3.27	רמת הכנסה
5.20	4.46	0.17	4.89	השכלת אב
5.27	4.31	0.20	4.96	השכלת אם
1.60	1.15	0.11	1.40	מס. היבחנות



## 4.2. מודלים דו-רמתיים

### 4.2.1. ניתוח שונות עם אפקטים מקריים

לוח 8 מציג את התוצאות של המודל:

#### לוח 8 – ניתוח שונות עם אפקטים מקריים

אפקט	מקדם	רכיב שונות	טעות תקן	ערך t/Z
קבוע	565.20	$\gamma_{00}$	1.0662	530(91)
מקרי	$\tau_{00}(u_{0j})$	101.37	15.5470	6.52
	$\sigma^2(r_{ij})$	10,015	24.2691	412.66

כפי שניתן לראות בלוח 8 הממוצע הכולל הוא 565.20 עם טעות תקן של 1.07, כלומר רווח בר סמך ברמה של 95% של ממוצע הממוצעים הוא  $(563.1, 567.3) = 565.2 \pm 1.96 * 1.07$ .

ערך זה (565.2) הינו בר השוואה לערך המתקבל בלוח 1 עבור רמה שנייה (565.17).

נזכיר שה-ICC (Interclass Correlation Coefficient) הינו מדד הן למידת ההומוגניות התוך-קבוצתית והן מדד להטרוגניות הבין-קבוצתית.

מתקבל עבור  $\tau_{00} = 101.37$  ו  $\sigma^2 = 10,015$ :

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{101}{101 + 10,015} = 0.01$$

המשמעות של ערך נמוך זה היא שרוב השונות נמצאת בתוך נוסח, או במילים אחרות הנבחן לא צריך לדאוג לאיזה נוסח הוא ניגש.

נשים לב שערכים אלו ( $\tau_{00}$  ו  $\sigma^2$ ) הינם שונויות. לכן, על מנת שיהיו ברי-השוואה לאלו בלוח 1, יש לחשב שורש ריבועי. מתקבל

$$\sqrt{\tau_{00}} = \sqrt{101.37} = 10.07 \quad \text{ו} \quad \sigma = \sqrt{\sigma^2} = \sqrt{10,015} = 100.07$$

ואכן ערכים אלה דומים מאוד לאלו בלוח 1.

רווח בר-סמך ברמה של 95% לממוצע של נוסח הוא  $(545.5, 584.9) = 565.2 \pm 1.96 * \sqrt{101}$ .

לבסוף נזכיר שמודל זה לא כולל חזאים והוא משמש בעיקר בסיס להשוואה עם המודלים הבאים.

## 4.2.2. רגרסיה של הממוצעים

מודל זה בוצע בשלושה שלבים. ראשית הוכנס כל אחד מהחזאים לבד, וזאת כדי לבדוק את ההשפעה של כל חזאי בפני עצמו. בשלב השני הוכנסו כל החזאים ביחד כדי לבדוק את התרומה שיש לכל חזאי מעבר לחזאים האחרים. כפי שנראה מיד, תרומתם של שני חזאים בשלב הזה לא הייתה מובהקת ולכן בשלב השלישי הוכנסו למודל רק החזאים שתרומתם הייתה מובהקת בשלב השני וזאת על מנת לקבל את המודל הסופי.

לוח 9 מציג את תוצאות השלב הראשון שכלל שישה מודלים של רגרסיה עם חזאי יחיד.

לוח 9 – תוצאות שש רגרסיות עם חזאי יחיד לניבוי ממוצע הציון הכולל

חזאי רמה 2	$\gamma_{00}$	$\gamma_{01}$	$\tau_{00}$	אחוז שונות מוסברת
מין	485.96***	144.08***	80.40***	20.71%
גיל	544.91***	0.94	101.76	0%
רמת הכנסה	667.33***	-31.23**	93.27***	8.01%
השכלת אב	439.53***	25.69***	83.89***	17.26%
השכלת אם	463.70***	20.47***	84.08***	17.08%
מספר היבחנות	525.44***	28.38**	91.42***	9.84%

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$

למרות שהעמודה השנייה ( $\gamma_{00}$ ) של החותך יוצאת כביכול מובהקת ברמה של 0.001, אין לכך משמעות היות ולא מרכזו את החזאים השונים, ולכן נתעלם ממנה. למרות שהשיפוע של משתנה הגיל לא מובהק (פרט לכך כל השיפועים מובהקים ברמה של לפחות 0.01), נכלול גם אותו במודל הרגרסיה שכולל את כל החזאים.

עבור שלב זה לא ננתח את השיפועים והשונויות; ההסברים יינתנו עבור המודל הסופי בלבד.

לוח 10 מציג את התוצאות של השלב השני שבו כל החזאים הוכנסו ביחד למודל.

**לוח 10 – תוצאות הרגרסיה של ממוצעי החזאים לניבוי ממוצע הציון הכולל – כל החזאים ביחד**

אחוז שונות מוסברת	$\tau_{00}$	$\gamma_{01}$	$\gamma_{00}$	חזאי רמה 2
		68.46**		מין
		-0.88		גיל
69.02%	31.40***	-40.15***	394.00***	רמת הכנסה
		27.59**		השכלת אב
		11.56		השכלת אם
		65.30***		היבחנות

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$

בדומה לתוצאות הרגרסיות שכללו חזאי יחיד והוצגו בלוח 9, גם במקרה זה החזאי גיל לא יוצא מובהק. נראה שאין תרומה ייחודית להשכלת אב, אם השכלת האב כלולה כבר במודל. בהתאם לכך, בשלב הבא, נוציא את שני החזאים גיל והשכלת אם ממודל הרגרסיה.

לוח 11 מציג את תוצאות השלב השלישי שבו נכללו ברגרסיה כל המשתנים שיצאו מובהקים בשלב הקודם (שזה הכול פרט לגיל ולהשכלת אם).

**לוח 11 – תוצאות הרגרסיה של ממוצעי החזאים לניבוי ממוצע הציון הכולל – המודל הסופי**

אחוז שונות מוסברת	$\tau_{00}$	$\gamma_{01}$	$\gamma_{00}$	חזאי רמה 2
		71.09***		מין
		-31.82***		רמת הכנסה
68.70%	31.73***	42.74***	332.91***	השכלת אב
		62.95***		מספר היבחנות

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$

הפעם לכל המשתנים תרומה מובהקת לניבוי ממוצע הציון הכולל (ברמה של 0.001), ולכן מודל זה יהווה עבורנו המודל הסופי.

השיפוע של מין יוצא חיובי (71.09) והמשמעות של תוצאה זו (נזכור שאנחנו קידדנו את המין כ-0 עבור זכר ו-1 עבור נקבה) היא שככל שיש יותר נשים בנוסח הציון הממוצע יותר גבוה. נשים לב, שתוצאה זו מנוגדת למה שאנו מצפים לו אינטואיטיבית מהנתונים שהוצגו בלוח 2.

השיפוע של רמת הכנסה הוא שלילי (-31.82) בהתאם לכיוון הסולם בשאלון (ראה לוח 3).

ככל שהשכלת האב ומספר ההיבחנות גבוהים יותר, כך ממוצע הציונים גבוה יותר (כפי שאפשר לצפות מהתבוננות בלוח 4 ובלוח 6 בהתאמה).

התקבל אחוז שונות מוסברת של 68.70%; נסביר כיצד התקבל ערך זה:

בלוח 8, שמהווה את הבסיס להשוואה, מתקבל ערך  $\tau_{00} = 101.37$ , ערך שהוא השונות בין ממוצעי הנוסחים, ללא כל המשתנים המסבירים. כאשר הוכנסו למודל המשתנים המסבירים, התקבל  $\tau_{00} = 31.73$ , ערך השווה לשונות בין ממוצעי הנוסחים במודל עם משתנים מסבירים (ברמה

$$\text{השנייה}). \text{ לכן אחוז השונות שהמודל הצליח להסביר הוא } 68.7\% = \frac{101.37 - 31.73}{101.37}$$

משום שהכנסנו רק משתנים מסבירים ברמה השנייה, מתקבל ערך של  $\sigma^2 = 10,015$  (לא מופיע בלוח), שהוא אותו הערך בדיוק כמו בלוח 8, כלומר ברמה הראשונה לא הצלחנו להסביר שום שונות, תוצאה שנראית סבירה לאור העובדה שלא הוכנסו למודל משתנים מנבאים ברמה זו בכלל.

בנוסף, קיבלנו גם:  $RMSE = \sqrt{31.73} = 5.6 \ll 10.3$ , כאשר הערך 10.3 הוצג בלוח 1. מכאן, כי יש שיפור משמעותי באיכות הניבוי. נראה כי יש לבחור במודל זה לצורך בקרת האיכות. אנו נתקף אותו בעזרת מדגם נוסף שהושאר בצד לצורך זה – מדגם התיקוף.

לצורך התיקוף נתעד את המשוואה שקיבלנו:

$$+ \text{ השכלת אב} * 42.74 + \text{ רמת הכנסה} * 31.82 - \text{ מין} * 71.09 + 332.91 = \text{ ממוצע נוסח מספר היבחנות} * 62.95$$

#### 4.2.3. רגרסיה עם מקדמים מקריים

בדומה למודל הקודם (הרגרסיה של הממוצעים) מקדמי המודל חושבו בשני שלבים (במקרה זה, כל המקדמים של החזאים אשר הוכנסו ביחד למודל היו מובהקים ולכן השלב השלישי היה מיותר).

לוח 12 מציג את תוצאות השלב הראשון שבו מדובר למעשה בשישה מודלים שכל אחד מהם כולל חזאי יחיד.

**לוח 12 – תוצאות שש רגרסיות עם מקדמים מקריים שבכל אחת מהן חזאי יחיד**

אחוז שונות מוסברת	$\sigma^2$	$\gamma_{10}(SD)$	$\gamma_{00}(SD)$	חזאי רמה 1
3.95%	9,619.50	-39.95(4.28)	587.15(10.36)	מין
0.41%	9,974.43	-0.89(1.66)	584.85(34.51)	גיל
7.92%	9,222.85	-26.32(1.92)	651.32(12.33)	רמת הכנסה
9.44%	9,070.51	20.24(1.03)	466.21(9.10)	השכלת אב
9.58%	9,056.02	21.42(2.48)	458.87(14.41)	השכלת אם
2.64%	9,751.23	22.58(4.03)	533.59(10.20)	מס. היבחנות

כל המקדמים וסטיות התקן היו מובהקים ברמה של 0.001.  
לוח 13 מציג את התוצאות של השלב השני בו הרגרסיה כללה את כל החזאים.

**לוח 13 – תוצאות רגרסיה עם מקדמים מקריים שכללה את כל החזאים**

אחוז שונות מוסברת	$\sigma^2$	$\gamma_{10}(SD)$	$\gamma_{00}(SD)$	חזאי רמה 1
		-37.25(4.61)		מין
		-0.88(1.20)		גיל
20.25%	7,987.19	-16.57(1.76)	533.39(29.35)	רמת הכנסה
		8.34(1.14)		השכלת אב
		11.19(1.39)		השכלת אם
		21.21(3.62)		מס. היבחנות

כל המקדמים וסטיות התקן היו מובהקים ברמה של 0.001.  
הפעם, היות והמודל כלל חזאים ברמה הראשונה, צפוי פְּתַח בשונות ברמה הראשונה המיוצגת על ידי  $\sigma^2$ . הערך שמתקבל עבור המודל הוא 7,987.19 ובהשוואה לערך של 10,015 (כפי שהתקבל במודל הריק מחזאים לוח 8).

$$\frac{10,015 - 7,987.19}{10.015} = 20.25\% \text{ נמצא שאחוז השונות המוסבר הוא } 20.25\%$$

נשים לב שערך זה נמוך משמעותית מן הערך שמתקבל במודל הרגרסיה של הממוצעים (68.7%). בנוסף, מתקבל:  $RMSE = \sqrt{7,987} = 89.37 \approx 100.53$ , כאשר הערך 100.53 לקוח מלוח 1. כלומר יכולת הניבוי של מודל זה חלשה יחסית. מסיבה זו וכן משום שאחוז השונות המוסברת במודל נמוך יחסית, בחרנו שלא לתקף אותו.

### 4.3. תיקוף

ראינו שמודל הרגרסיה של הממוצעים מצליח לנבא 69% מן השונות של הציון הממוצע בנוסח, עם RMSE השווה ל-5.6 (כאשר טווח הציונים בנוסח נע בין 532 ו-586 עם סטיית תקן של 10.3). לעומת זאת נמצא שמודל הרגרסיה עם מקדמים מקריים מצליח לנבא רק 20% מן השונות של ציוני הנבחנים עם RMSE השווה ל-89.37 (כאשר טווח הציונים עבור נבחן בודד נע בין 227 ו-800 עם סטיית תקן של 100).

בהתאם לכך תוקף רק מודל הרגרסיה של הממוצעים (המודל השני). התיקוף בוצע על מדגם נוסף שכלל 34 נוסחים שלא נכללו במדגם שעליו חושבו הפרמטרים (מדגם התיקוף).

לוח 14 מציג את נתוני הניבוי עבור 34 הנוסחים: מספר נוסח, הממוצע המנובא, הממוצע בפועל וההפרש ביניהם.

**לוח 14 – מספר נוסח, ממוצע מנובא, ממוצע בפועל והפער ביניהם עבור 34 נוסחים במדגם התיקוף**

# נוסח	מנובא	אמיתי	מנובא-אמיתי
1	570.08	569.93	0.15
2	573.85	573.71	0.14
3	576.34	578.68	-2.35
4	567.45	560.11	7.34
5	562.19	566.28	-4.09
6	572.27	571.33	0.93
7	575.70	571.84	3.87
8	579.13	573.70	5.44
9	557.98	565.65	-7.67
10	564.00	567.94	-3.95
11	560.97	565.19	-4.21
12	579.44	579.51	-0.07
13	567.67	574.42	-6.75
14	572.57	569.44	3.13
15	579.24	570.59	8.65
16	555.55	575.01	-19.46
17	569.89	578.52	-8.63

9.86	561.49	571.35	18
7.25	554.26	561.51	19
0.87	552.22	553.09	20
8.14	561.28	569.43	21
7.37	562.11	569.49	22
4.48	561.48	565.96	23
8.02	552.59	560.61	24
-4.26	559.66	555.39	25
-3.06	572.21	569.15	26
-8.89	555.11	546.23	27
-1.46	574.42	572.96	28
1.82	564.56	566.38	29
9.65	562.94	572.59	30
8.85	547.40	556.25	31
-2.25	554.10	551.85	32
1.21	571.41	572.62	33
7.72	564.25	571.97	34
<b>0.82</b>	<b>565.98</b>	<b>566.80</b>	<b>ממוצע</b>
6.60	8.24	8.32	ס.ת.
-19.46	547.4	546.23	מינימום
9.86	579.51	579.44	מקסימום

לפי המודל נִצְפָּה ש-32 תצפיות המהוות כ-95% מהממוצעים המנובאים ( $34 \cdot 0.95 = 32.2$ ) ייפלו בטווח של  $1.96 \pm$  סטיות תקן (5.6) מהממוצע האמיתי, כלומר רווח של עד 11 נקודות לכל כיוון. הבדיקה אכן מראה כי רק תצפית אחת (תצפית 16) חורגת מטווח זה. ממוצע הטעות על 34 הנוסחים שנכללו במדגם התיקוף הוא 0.82 נקודות בלבד (כאשר הממוצע של הפערים המוחלטים הוא 5.35).

## 5. דיון

מחקר זה השתמש בניתוח רב-רמתי כדי לחקור את הקשר בין משתני הרקע של נבחנים לבין הציון הכללי בבחינה הפסיכומטרית, וזאת על מנת לבצע בקרת איכות על חישוב הציונים של נוסח חדש.

הקשר נבדק בשתי רמות: ברמת הנבחן וברמת הנוסח.

- ברמת הנבחן הבודד נמצא אומנם קשר משמעותי בין משתני הרקע לציון (הקשר היה שונה בעבור נוסחים שונים), אך יחד עם זאת היכולת לנבא ציון לנבחן בודד הייתה נמוכה.
- ברמת הנוסח נמצא קשר חזק בין משתני הרקע המקובצים לממוצע הציון בנוסח וכן יכולת ניבוי גבוהה.

המודל עם יכולת הניבוי הטובה ביותר תוקף על בסיס נתונים שלא נכלל במדגם האמידה ונמצא תקף.

תוצאות המחקר תומכות בשימוש בשיטה זו לצורך בקרת איכות של תהליך חישוב הציונים בבחינה הפסיכומטרית. אפשר להשתמש בכלי זה באופן מיידי, טרם דיווח הציונים.

מומלץ להמשיך את המחקר על מנת לבדוק אם ניתן לשפר את הניבוי (ראה הצעות למחקרי המשך) וכן לוודא כי הקשר בין המשתנים הדמוגרפיים לציון בבחינה נשאר יציב.

התוצאות המעודדות של המחקר ממחישות את חשיבות בקרת האיכות ככלי שנועד לשפר את מידת הדיוק של מבחנים סטנדרטיים. בקרת איכות שוטפת מעלה את רמתו המקצועית של המבחן ומבטאת את אחריותו האתית של הארגון שעורך את המבחנים אל מול הנבחנים והמשתמשים בציון.

### 5.1. אחרית דבר

בזמן כתיבת שורות אלו הגיעו לידינו נתוני שנת 2014. משוואת המודל (זו המופיעה בפרק 4.2.2) יושמה על שנה זו (לאחר ביצוע סינון דומה לזה שנעשה במודל המקורי). נמצא שבאחד מתוך 8 נוסחים התוצאה הייתה גבולית (טעות ניבוי של 10.9), בעוד שעבור שאר הנוסחים הטעות הייתה קטנה. כפי שהוסבר בפרק התיקוף (פרק 4.3), צפוי שיעור מסוים של נוסחים שעבורם תתקבל סטייה מסוימת מהפער הצפוי על פי המודל. התבנית שהתקבלה עבור נוסחי 2014 תומכת בנכונות המודל.

בנוסף, נעשה ניסיון לאפיין את הערכים החריגים ולנסות ולהסביר ממה הם נובעים. לא נמצא הסבר לחריגות גם כאשר הנוסחים חולקו לפי שנה וגם כאשר בוצעה חלוקה לפי חודש הבחינה.



## 5.2. הצעות למחקרי המשך

מחקר זה השתמש בשישה משתני רקע דמוגרפים. אפשר לשקול שימוש במשתנים נוספים כמו:

- מועד הבחינה (או באופן אלטרנטיבי לעשות ניתוח נפרד לכל חודש)
- בגרות (דיווח עצמי). בעיה שכרוכה בשימוש בבגרות היא שכמחצית מהנבחנים חסרים נתון זה ובסה"כ רק כ-40% מהנבחנים יהיו בעלי נתונים מלאים אם יוחלט לכלול את המשתנה במודל.

אפשר לנסות ולהשלים נתונים חסרים בשיטות שונות (למשל בשיטת הזקיפה המרובה – multiple imputation) ולראות כיצד הניבוי יתנהג.

בעבודה זו התבצע ניבוי של הציון הכללי בבחינה. מומלץ גם לבנות מודל עבור כל אחד משלושת תחומי הבחינה: כמותי, מילולי ואנגלית, ולבדוק את טיב הניבוי של כל אחד מהתחומים. סביר להניח שמודל מסוג זה יהיה מדויק יותר, שכן כל משתנה דמוגרפי עשוי להשפיע אחרת על כל תחום בבחינה.

- סער י. ואורן כ. (2014). דוח סטטיסטי לשנים 1999-2013. ירושלים: מרכז ארצי לבחינות ולהערכה.
- Allalouf, A. (2007). Quality control procedures in the scoring, equating and reporting of test scores. *Educational Measurement: Issues and Practice*, 26, 36-46.
- Gamoran, A. Porter, C. A., Smithson, J. & White, P. A. (1997). Upgrading High School Mathematics Instruction: Improving Learning Opportunities for Low-Achieving, Low Income Youth. *Educational Evaluation and Policy Analysis*, Vol. 19, No.4, 325-338.
- Kidwell, E. R., Mossholder, W. K., Bennet, N. (1997). Cohesiveness and Organization Citizenship Behavior: A Multilevel Analysis Using Work Groups and Individuals. *Journal of Management*, Vol. 23, No. 6, 775-793.
- Kolen, M. J., Brennan, R. L. (2014). *Test Equating, Scaling, and linking*. Springer Series in Statistics 3<sup>rd</sup> ed. 2014.
- Li, D., Li, S., & von Davier, A. A. (2011). Applying time-serious analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York, NY: Springer-Verlag.
- Liu, M., Lee, Y.-H, & von Davier, A. A. (2012). *Detection of unusual administrations using a linear mixed effects model*. Paper presented at the international meeting of the Psychometric Society, Lincoln, NE.
- Montgomery, D. C. (2009). *Introduction to statistical quality control, sixth edition*. Hoboken, NJ: John Wiley & Sons.
- Rumberger, W. R. (1995). Dropping out of Middle School: A multilevel Analysis of Students and Schools. *American Educational Research Journal*, Vol. 32, No 3, 583-625.
- Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand.
- The College Board (2013). Total Group Profile Report.  
<http://media.collegeboard.com/digitalServices/pdf/research/2013/TotalGroup-2013.pdf>
- Wang, J., Xie, H. & Fisher, J.H. (2011). *Multilevel Models. Applications using SAS*.
- Wei, Y. (2013). *Monitoring TOEIC Listening and Reading Test Performance Across Administrations Using Examinees' Background Information*. The Research Foundation for the TOEIC Tests: A Compendium of Studies: Volume II. Princeton, NJ: Educational Testing Service, Sep 2013, p11.1-11.28.