

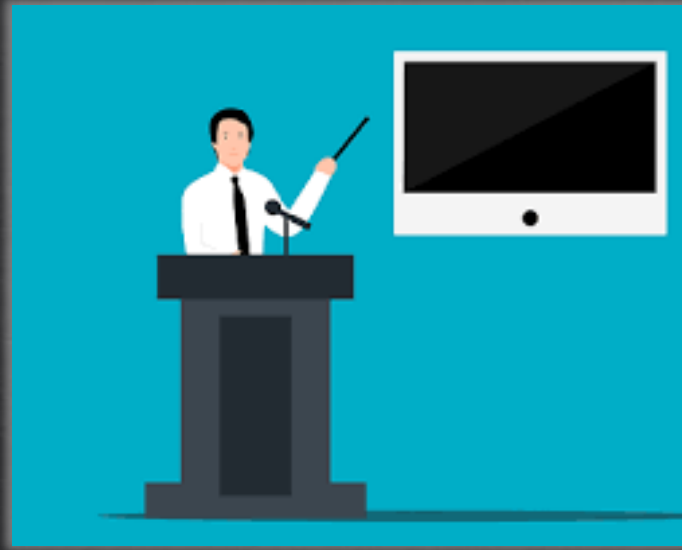
בדיקת תלויות בתוך מקבץ פריטים בגישת IRT

צור קרליץ

נעם כאהן



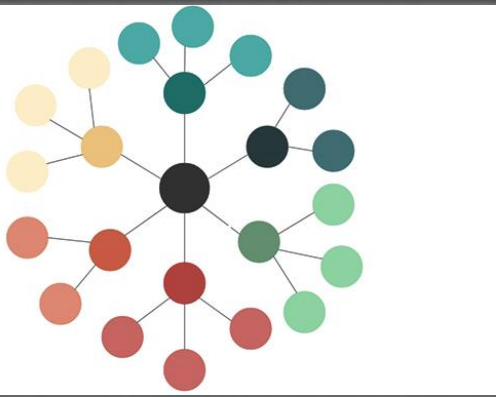
מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتقييم
מיסודן של האוניברסיטאות בישראל



מבנה ההרצאה

- ◇ מהו מקבץ פריטים?
- ◇ IRT על קצה המזלג
- ◇ מקבצי פריטים והנחת אי-תלות מקומית ב IRT
- ◇ גישות לבדיקת תלות במקבצי פריטים
 - ◇ מדד Q_3
 - ◇ מודל Testlet
- ◇ דוגמאות מהדמיית נתונים, מהבחינה הפסיכומטרית ומהמיצ"ב
- ◇ תודה לרשות הארצית למדידה והערכה בחינוך (ראמ"ה) על האישור להשתמש בנתוני המיצ"ב
- ◇ סיכום ומסקנות

מהו מקבץ פריטים?



◇ מקבץ הוא אוסף של פריטים המבוססים על גריין אחד (Wainer & Kiely, 1987)

◇ לדוגמה, פרטי הבנת הנקרא, פריטי הסקה מתרשים / טבלה, פריטים מרובי שלבים

◇ שמות אחרים: פריט אב + שאלות, אשכול פריטים (Item bundle), פרקון, טסלט (testlet)

◇ למקבצי פריטים יש יתרונות פרקטיים:

◇ דיאגנוסטיקה: מאפשרים לבדוק הבנה וידע "לעומק"

◇ חסכוניות: הזמן שהנבחן משקיע בהבנת גריין אחד משמש לו לפתרון כל השאלות

◇ רלוונטיות: פעמים רבות בעולם האמיתי אנו נדרשים לפתור מספר בעיות הקשורות לנושא אחד

תלות במקבץ פריטים

חשיבה מילולית - פרק ראשון

- 7 -

מועד פברואר 2011

הבנת הנקרא (שאלות 25-30)

קראו בעיון את הקטע, וענו על השאלות שאחריו.

(1) כשאנו מעוניינים לדעת מהי הגדרתו של מושג מסוים, אנו פונים למילון. מאז תקופת יוון העתיקה ועד ימינו עסקו פילוסופים בשאלה באילו תנאים צריכה הגדרה לעמוד כדי שתיחשב הגדרה טובה.

(5) על פי גישת האסנציאליזם שהתפתחה במאה ה-19, הגדרת מושג צריכה לכלול את כל התכונות ההכרחיות והמספיקות על מנת להיכלל באותו מושג. כך, בהיתקלנו במקרה פרטי מסוים נוכל לפסוק אם הוא שייך לאותו מושג, וזאת על ידי השוואת המקרה הפרטי להגדרה תוך בדיקה אם הוא ניחן בכל אותן תכונות. למשל, הגדרת המושג "מספר ראשוני" - "מספר שלם שאינו מתחלק בלי שארית בשום מספר שלם מלבד בעצמו ובאחת" - מאפשרת לקבוע בנוגע לכל מספר אם הוא נכלל בהגדרה או לא.

(10) פילוסופים שונים ביקרו את גישת האסנציאליזם, וטענו כי השאיפה להגדיר מושגים באופן זה היא יומרנית, שכן אי-אפשר להחיל אותה על רוב המושגים השגורים בפנינו. לדוויג ויטגנשטיין, פילוסוף אוסטרי שפעל בתחילת המאה ה-20, עמד על שתי בעיות עיקריות בהגדרת מושגים: הבעיה הראשונה כותה בפיו "גבולות מעורפלים".

השאלות

12. מתוך זוגות הכיוונים הבאים, בין אילו שני כיוונים הפרש בעוצמת הרוח הממוצעת הוא הגדול ביותר?

- (1) צפון ודרום
- (2) מזרח ומערב
- (3) צפון-מזרח ודרום-מערב
- (4) צפון-מערב ודרום-מזרח

13. ידוע שעוצמת הרוח בזמן נתון יכולה להיות בין מחצית לבין כפליים העוצמה הממוצעת. איזו מהטענות הבאות אינה נכונה?

- (1) עוצמת הרוח מכיוון צפון-מערב יכולה להיות 85 קמ"ש
- (2) עוצמת הרוח מכיוון דרום יכולה להיות 20 קמ"ש
- (3) עוצמת הרוח מכיוון מזרח יכולה להיות 10 קמ"ש
- (4) עוצמת הרוח מכיוון מערב יכולה להיות 20 קמ"ש

14. באי גולו-גולו שכוחות הרוחות ה_____ היא הגדולה ביותר.

- (1) צפונית
- (2) מערבית
- (3) דרומית
- (4) מזרחית

15. אילו לוחות הדרומיות הייתה שכוחות שווה, מה היה הממוצע של עוצמות הרוחות הדרומיות (בקמ"ש)?

- (1) 10
- (2) 20
- (3) 15
- (4) 25

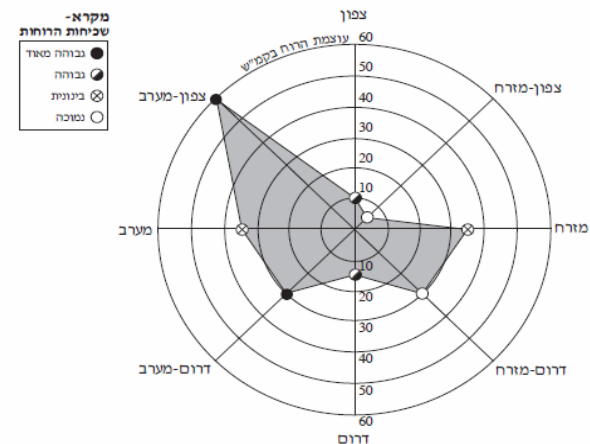
הסקה מתרשים (שאלות 12-15)

עיינו היטב בתרשים שלפניכם, וענו על ארבע השאלות שאחריו.

בתרשים מתוארות הרוחות באי גולו-גולו. כל אחת מהנקודות המסומנות בתרשים מייצגת רוח מכיוון מסוים, וכן את עוצמתה הממוצעת (בקמ"ש) ואת שכוחותה. עוצמת הרוח מכל כיוון מיוצגת על-ידי מרחק הנקודה ממרכז התרשים, ושכוחות הרוח מיוצגת על-ידי צורת הנקודה (ראו מקרא).

נגדו: רוחות צפוניות הן הרוחות מכיוון צפון, צפון-מזרח וצפון-מערב. רוחות מערביות הן הרוחות מכיוון מערב, צפון-מערב ודרום-מערב. באופן דומה מוגדרות רוחות דרומיות ורוחות מזרחיות.

לדוגמה: מכיוון דרום-מזרח עוצמת הרוח הממוצעת היא 30 קמ"ש, ושכוחותה נמוכה.



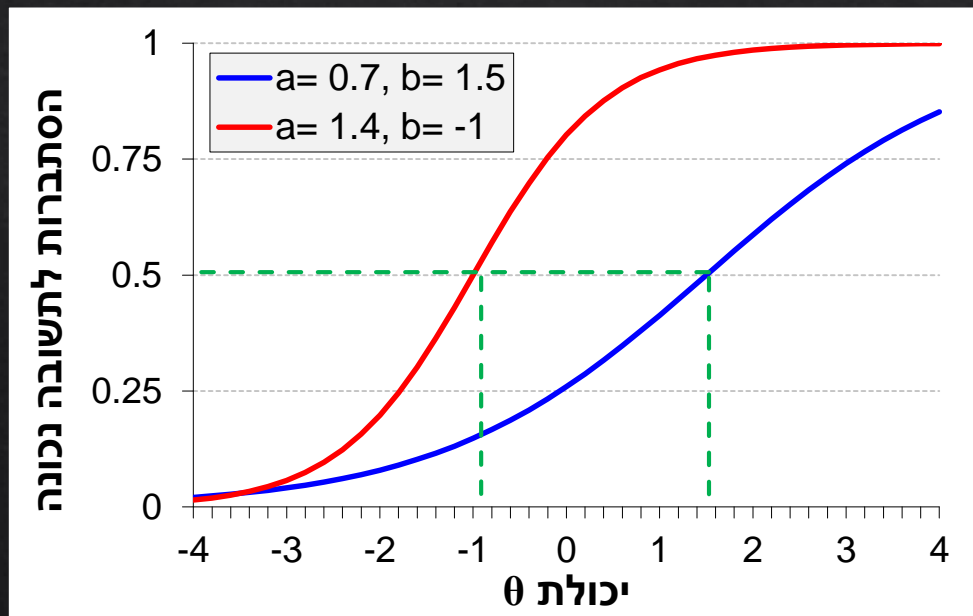
שימו לב: בתשובתכם לכל שאלה התעלמו מנתונים המופיעים בשאלות האחרות.

ידע מוקדם על נושא הטקסט יכול לסייע בפתרון כל השאלות במקבץ, אבל לא יסייע בפתרון השאלות האחרות בפרק. כלומר, יש תלות רק בין הפריטים במקבץ.

המידה בה נבחנים מבינים את התרשים משפיעה על יכולתם לפתור נכון כל אחד מהפריטים ולכן יוצרת תלות ביניהם. הכרות מוקדמת עם סוג התרשים יכולה לעזור בפתרון נכון של הפריטים.

תורת התגובה לפריט (IRT) Item Response Theory

- ◇ מודלים של IRT מקשרים בין יכולת חבויה ורציפה (θ) לבין הסתברות לתשובה נכונה בפריט במבחן.
- ◇ הקשר בין ביצוע בפריט לרמת היכולת של הנבחן מתואר בעזרת אופיין פריט (ICC)
- ◇ הסולם המשמש לתיאור יכולות הנבחנים (סולם לוג'יט) משמש גם לתיאור קושי הפריטים.
- ◇ קושי הפריט (b) הוא היכולת עברה ההסתברות לענות על הפריט היא 0.5.
- ◇ פרמטר ההבחנה (a) מתאר את שיפוע האופיין – עוצמת ההבחנה בין נבחן חזק לחלש.



אופיין הפריט Item Characteristic Curve

- ◇ מודל 1PL (Rasch) – רק פרמטר של קושי
- ◇ מודל 2PL – פרמטרים של קושי והבחנה
- ◇ מודל 3PL – פרמטרים של קושי, הבחנה וניחוש (c)

$$P(x = 1) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}$$

הסתברות לתשובה נכונה בפריט
 הבחנה
 יכולת הנבחן
 קושי הפריט

מודל 2PL



אי תלות מקומית ב- IRT

◇ הנחת הבסיס במודלים של IRT היא קיומה של אי-תלות מקומית.

◇ בהינתן רמת יכולת קבועה, התגובות לפריטים הן בלתי תלויות.

◇ אם נחזיק את היכולת קבועה (נבטל את השפעתה) לא נראה קשר בין הביצוע בפריטים.

◇ משמע, היכולת היא הגורם היחידי המקשר בין הביצוע בפריטים שונים (כלומר, המבחן חד ממדי)

◇ אם ההנחה מופרת, סימן שיש גורמים נוספים, לא רלבנטיים, שמשפיעים על הביצוע במבחן.

◇ התעלמות מהתלות בין פריטים מובילה לבעיות באמידת המודל:

◇ הטייה באומדני היכולת

◇ הטיית באומדני מאפייני הפריטים

◇ הערכת חסר של טעויות התקן של האומדנים

◇ הערכת יתר של האינפורמציה של המבחן וניפוח המהימנות

◇ Thissen, Steinberg & Mooney, 1989; Yen, 1993; Tuerlincks & De Boeck, 2001

◇ **שאלת המחקר:**

האם קיימת תלות מקומית במבחנים הכוללים מקבצי פריטים?

מדד Q_3 לבדיקת אי-תלות מקומית ברמת הפריטים

◇ Yen (1984) הציעה את המדד Q_3 המבוסס על המתאם בין השאריות של זוגות פריטים.

◇ עבור כל פריט, השארית היא הפרש בין הציון (0 או 1) לניבוי (הסתברות לתשובה נכונה): $d_{is} = x_{is} - P_i(\theta_s)$

◇ כלומר, שארית = הביצוע בפריט בניכוי החלק שמוסבר על ידי היכולת.

◇ המדד Q_{3ij} הוא המתאם בין השאריות של פריט i ופריט j , מעבר לכל הנבחנים.

◇ אם היכולת מסבירה את כל הקשר בין הפריטים, אז המתאם בין השאריות יהיה קרוב לאפס, כלומר מתקיימת הנחת אי-תלות מקומית בין הפריטים.

◇ אם יש מתאם בין השאריות אז יש גורם נוסף שמקשר בין הפריטים מלבד היכולת.

◇ ערך Q_3 גבוה מ-|0.2| מצביע על קיומה של תלות-מקומית בין זוג הפריטים.

◇ למדד יש הטייה קטנה וכדי לבטלה נהוג להחסיר את הממוצע של מדדי ה- Q_3 של כל הפריטים.

◇ גישה אחת לבדיקת התלות במקבץ היא לבחון את הגודל של Q_3 בין הפריטים בתוך המקבץ לעומת הפריטים מחוץ למקבץ.

הדגמה של Q_3

◇ דוגמה 1 – מיצ"ב

◇ 67 פריטי הבנת הנקרא מנוסחי מיצ"ב עברית לכיתות ה' בין 2016 ל-2018.

◇ 3 נוסחים, 2 קטעי הבנת הנקרא בנוסח, כ- 11 פריטים במקבץ.

◇ N כולל = 59,513 תלמידים, כ- 16,500 לנוסח.

◇ חלק מהפריטים היו פוליטומיים (מרובי ציונים), והפכנו אותם לדיכוטומיים לצורך הניתוח.

◇ דוגמה 2 – הבחינה הפסיכומטרית

◇ 12 פרקי חשיבה כמותית (20 שאלות בפרק, תרשים אחד בפרק, 4 שאלות במקבץ)

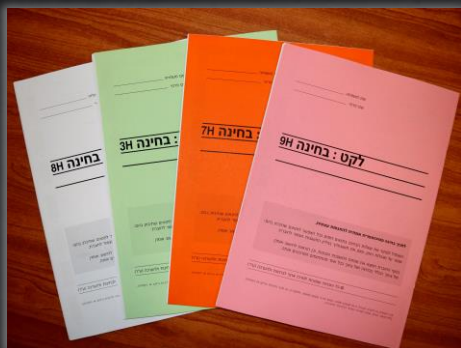
◇ 11 פרקי חשיבה מילולית (23 שאלות בפרק, קטע הבנת הנקרא אחד, 6 שאלות במקבץ)

◇ 11 פרקי אנגלית (22 שאלות בפרק, 2 קטעי הבנת הנקרא, 5 שאלות בכל מקבץ)

◇ N כולל = 275,964 נבחנים, כ- 5,200 לפרק.



מדדי יעילות וצמיחה בית-ספרית



מדדי Q₃ בשני המדגמים

נתוני המיצ"ב (מודל 1PL):

מדדים של Q₃ בין זוגות כל הפריטים

ממוצע	ס.ת.	מינימום	מקסימום	# זוגות Q ₃ >0.2
-0.05	0.04	-0.14	0.59	1 מתוך 2,211

כל המתאמים קטנים מ-0.2

מסקנה: אין עדות לתלות בין הפריטים בתוך המקבצים.

ממוצע Q₃ בתוך מקבץ ובין מקבצים

מקבץ	1	2	3	4	5	6
1	-0.03					
2	-0.06	-0.05				
3			-0.05			
4			-0.07	-0.02		
5					-0.04	
6						-0.01

נתוני הבחינה הפסיכומטרית (מודל 2PL):

מדדים של Q₃ בין זוגות פריטים, מעבר לכל הפרקים

ממוצע	ס.ת.	מינימום	מקסימום	# זוגות Q ₃ >0.2	
-0.04	0.04	-0.16	0.18	0 מתוך 2,280	כמותי
-0.03	0.04	-0.20	0.22	4 מתוך 2,783	מילולי
-0.04	0.03	-0.16	0.10	0 מתוך 2,541	אנגלית

מסקנה: אין עדות לתלות בין הפריטים בתוך המקבצים.

ממוצע Q₃ (וס.ת.) בתוך ובין מקבצים ופריטים חופשיים

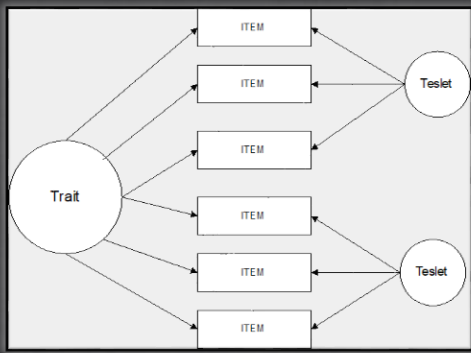
כמותי	מילולי	אנגלית	
0.04 (0.03)	0.00 (0.05)	-0.01 (0.02)	בתוך מקבץ
-0.06 (0.01)	-0.02 (-)	-0.06 (0.01)	בין מקבצים
-0.03 (0.01)	-0.02 (0.02)	-0.03 (0.01)	בין פריטים חופשיים
-0.06 (0.01)	-0.06 (0.01)	-0.05 (0.01)	בין פריטים למקבצים



גישות למידול מקבצי פריטים ב IRT

- ◇ מידול של כל המקבץ כ"סופר-פריט" מסוג ניקוד חלקי (Sireci, Thissen, & Wainer, 1991)
- ◇ במקום 3 פריטים נקבל פריט אחד עם 4 רמות ניקוד (0, 1, 2, 3)
- ◇ פותר את בעיית התלות בין הפריטים במקבץ אבל מאבד מידע דיאגנוסטי חשוב (תבנית התגובה לפריטים)
- ◇ התייחסות לתלות בתור ממד נוסף ושימוש במודל רב-ממדי (MIRT; Reckase, 1997)
- ◇ אפשרי אם ניתן לאפיין את הממד הנוסף שמשפיע על הביצוע במקבץ
- ◇ מודלים רב-ממדיים הם מסובכים וקשים יותר לפירוש
- ◇ מידול האפקט של כל מקבץ כפרמטר נוסף במודל IRT למקבצי פריטים
- ◇ Testlet Model (Wainer & Kiely, 1987; Bradlow, Wainer, & Wang, 1999)
- ◇ הפרמטר מייצג את השינוי בהסתברות לתשובה נכונה שנובע מהמפגש בין נבחן למקבץ מסוים.

מודל טסטלט



$$P(X_{ij} = 1) = \frac{e^{\alpha_j(\theta_i + \gamma_{id(j)} - \beta_j)}}{1 + e^{\alpha_j(\theta_i + \gamma_{id(j)} - \beta_j)}}$$

◇ θ היא יכולת נבחן i , α הוא השיפוע של אופיין פריט j , β הוא הקושי של פריט j .

◇ יש גם וריאציה עם פרמטר לניחוש, וריאציה רב-ממדית, ועוד...

◇ γ מייצג את האינטראקציה בין מקבץ $d(j)$ (זה שמכיל את פריט j) והנבחן i .

◇ הפרמטר מייצג את הסטייה האקראית ביכולת הנבחן בעקבות קיומו של ממד משני הקשור למקבץ

◇ $\gamma_{id(j)} \sim N(0, \sigma^2_{d(j)})$ – הפרמטר מתפלג נורמלית עם ממוצע 0 ושונות כלשהי.

◇ **אפקט הטסטלט** – $\sigma^2_{d(j)}$ = השונות של פרמטרי γ בתוך כל מקבץ, מעבר לכל הנבחנים.

◇ אם לפרמטר אין שונות, אז האינטראקציה בין המקבץ לנבחן היא לכל הנבחנים- אין תלות מקומית.

◇ גודל השונות מייצג את מידת התלות בתוך המקבץ. (Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000))

◇ שונות קטנה מ- 0.25 היא זניחה- אינה משפיעה על הפרמטרים האחרים במודל.

◇ שונות בין 0.25 ל- 1 אינה זניחה אבל אינה חמורה.

◇ שונות מעל 1 משפיעה על הפרמטרים האחרים בצורה משמעותית.

RMSE = שורש ממוצע הפרשים מרובעים

Testlet	2PL	
0.999	0.997	מתאם עם b
0.003	0.095	עם RMSE b

הדגמה - נתונים מסומלצים

יצרנו נתונים על פי מודל הטסטלט:

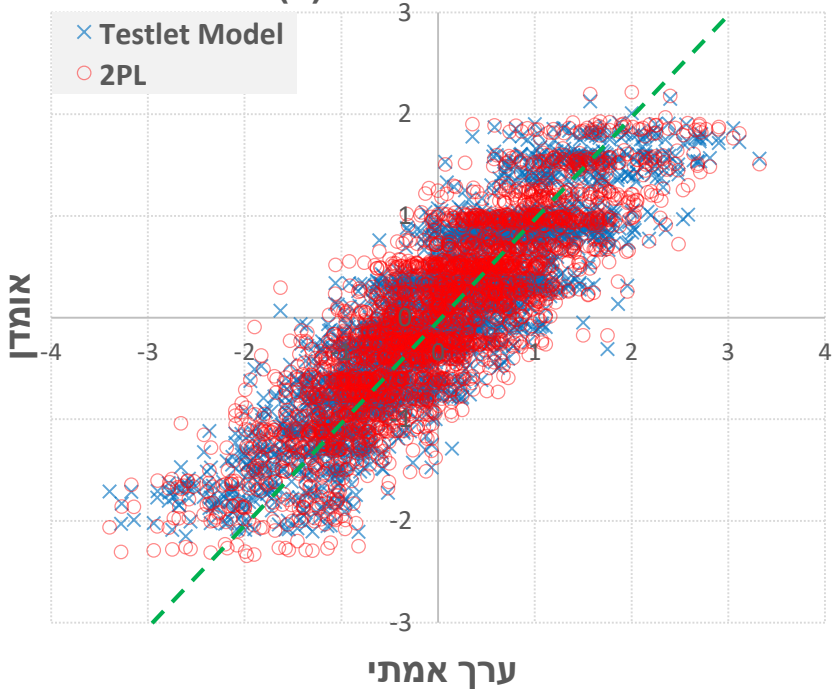
3 מקבצים (שונות אפקט הטסטלט = 0, 0.5, 1.6)

בכל מקבץ 4 פריטים: (קל, קשה) x (מבחין, פחות מבחין)

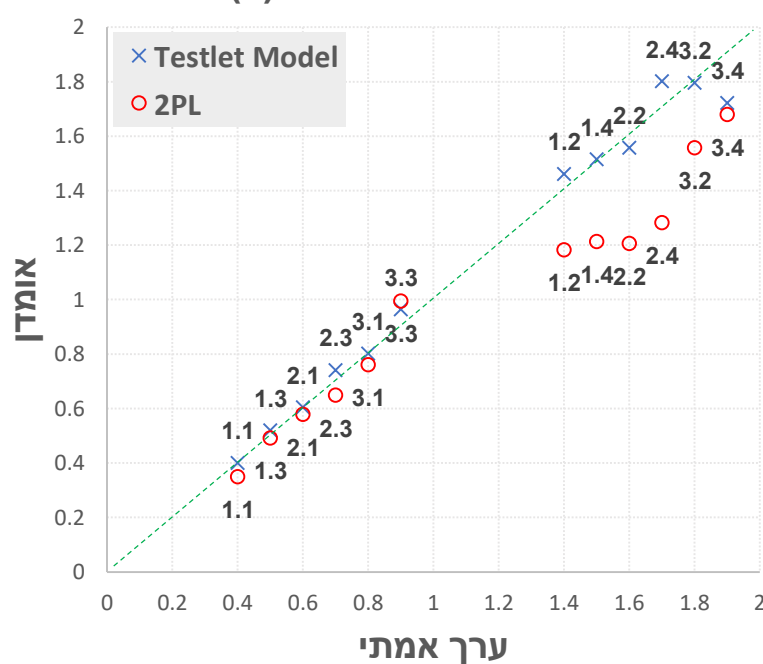
התפלגות הנבחנים – נורמלית סטנדרטית

אמדנו מודל 2PL ומודל טסטלט והשוונו את אומדני הפרמטרים לערכם האמתי

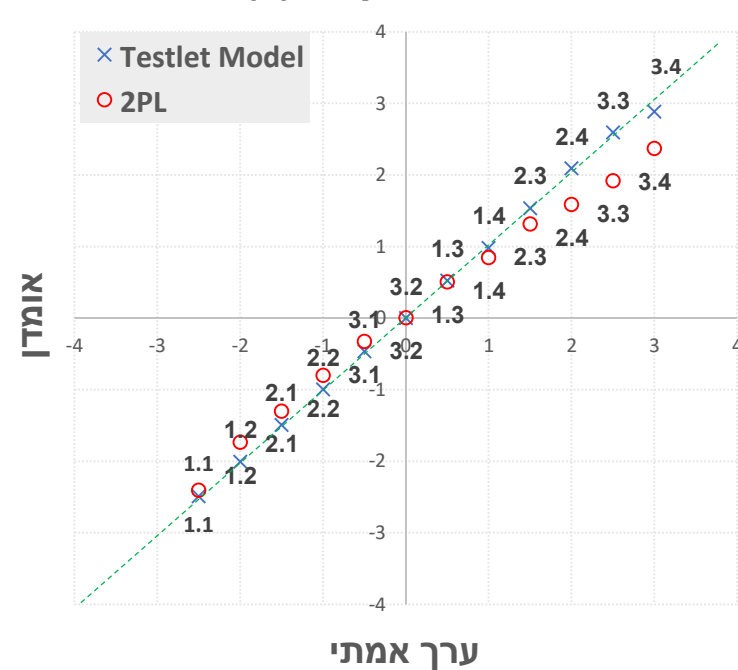
אומדני יכולת (θ)



פרמטר הבחנה (a)



פרמטר קושי (b)



שימוש במודל טסטלט בשני המדגמים

◆ נתוני המיצ"ב:

מקבץ	אפקט טסטלט $\sigma^2_{d(j)}$
1	0.03
2	0.03
3	0.01
4	0.17
5	0.10
6	0.09

◆ אומדנים של שונות אפקט הטסטלט במקבצים שונים:

◆ מסקנה: כל השונויות זניחות (קטנות מ-0.25). אין עדות לאפקט מקבץ.

◆ נתוני הבחינה הפסיכומטרית:

◆ סטטיסטיקה תיאורית של שונות אפקט הטסטלט בכל תחום:

תחום	# מקבצים	אפקט ממוצע	ס.ת	אפקט מקסימלי	% $0.25 < t < 1$	אחוז מעל 1
כמותי	24	0.35	0.46	1.61	30%	12.5%
מילולי	23	0.26	0.23	0.68	43%	0%
אנגלית	33	0.16	0.10	0.39	21%	0%
סה"כ	80	0.25	0.29	1.61	30%	3.75%

◆ מסקנה:

◆ במרבית המקרים האפקט הוא זניח.

רק בכמותי ישנם מעט מקרים בהם האפקט הוא משמעותי.

סיכום ומסקנות

- ◇ מקבצי פריטים עשויים ליצור מצב בו הביצוע בפריטים תלוי בגורמים שאינם קשורים ליכולת הנמדדת אלא קשורים לגריין המשותף לפריטים במקבץ.
- ◇ עשוי להיות בעייתי בניתוח נתוני המבחן, במיוחד עם מודלים של IRT המניחים אי-תלות מקומית בין הפריטים.
- ◇ אפשר להשמיט פריטים שמצביעים על תלות, להפוך מספר פריטים ל"סופר פריט" או למדל את קיום המקבץ.
- ◇ הצגנו שתי גישות לבדיקת מידת התלות בין הפריטים בתוך מקבץ:
 - ◇ מדד Q3 הבודק את המתאם בין שאריות הפריטים (ביצוע בפועל בניכוי ביצוע מנובא)
 - ◇ מודל טסטלט האומד את שונות הפרמטר שמייצג אינטרקציה יחודית בין נבחן למקבץ
- ◇ הדגמנו את השימוש בשתי הגישות על נתונים מסומלצים ונתוני אמת
 - ◇ בסימולציות, כשקיימת תלות במידה משמעותית בין פרטי המקבץ, נצפתה ירידה בדיוק בעיקר באומדני הפריטים.
 - ◇ במרבית המקרים בנתוני האמת לא נמצאה תלות גדולה בין הפריטים במקבץ.
 - ◇ שתי הגישות לא בהכרח מסכימות בינהן לגבי קיומה של תלות או עוצמתה.
 - ◇ בקבלת ההחלטה לגבי אופן הטיפול במקבץ על בסיס המדדים הרלבנטיים יש להתייחס גם לתוכן הפריטים.

תודה על ההקשבה

