פיתוח מדד יחיד לאיכות מעריכים של מבחני ביצוע

אבי אללוף, דביר קלפר, האשם ניג'ם, ג'ני שמולביץ

25.1.2016

מבוא 1

שמירה על מהימנות ואובייקטיביות של ציונים על מבחני ביצוע דורשת מחוון ברור ומעריכים איכותיים. **מיהו מעריך איכותי?**

מעריך שהוא בעל התכונות הבאות:

- תורם בסדנאות – **משתתף פעיל ובאווירה חיובית**
- מדויק – **מעריך לפי המחוון**
- מהיר – **בעל הספק גבוה. משמעות כלכלית**
- זמין – **מתמיד, מוכן להעריך כשמתבקש**
- יציב לאורך זמן – **שומר על תכונותיו (החיוביות) לאורך זמן.**

המדדים שתוארו לעיל מאפשרים בקרה על טיב המעריכים, אך חסר **מדד יחיד, הלוקח בחשבון מספר מדדים.**

מבוא 2

שאלת המחקר: כיצד ניתן לחשב מדד יחיד לאיכות מעריכים של מבחני ביצוע בהתבסס על מספר מדדים שונים?

לשאלה יש שני היבטים:

תיאורטי – כיצד מחשבים מדד יחיד (למשל, מה המשקלות של המדדים השונים)

מעשי – השימוש במדד היחיד יגדיל את מהימנות ההערכה (ואולי לחסוך במשאבים)

מאל"ו מעסיק מאות מעריכים במבחני ביצוע, ועליו לבקר את איכותם.

במטלת הכתיבה מועסקים, מדי שנה, כ-250 מעריכים (עברית, ערבית, 9 שפות נוספות).

במבחנים אחרים (חיבור של בחינת ידע בעברית, שאלוני: מור, מרקם, מרב, אבני ראשה; בחינת מתאם) עוד כ-100.

לפני השיטה....

משתנה חשוב שנחקר כאן הוא הספק (חיבורים ליחידת זמן)
בספרות אין התייחסות למשתנה זה

מחקר קודם שנערך על מעריכי השאלונים של מור מרקם לא
מצא קשר מובהק בין ההספק לאיכות

ניתן, כמובן, לערוך "שיפוט קליני" ולתת משקלים לפי תפיסתנו
על חשיבות המדדים; המחקר מהווה "שיפוט סטטיסטי"

~~ מבחן ~~

מטלת הכתיבה המהווה חלק מהתחום המילולי בבחינה הפסיכומטרית וכוללת גם חלקים מילולי וכמותי.

נבחנים מקבלים גריין קצר. הזמן המוקצב: 30-35 דקות. החיבור מוערך על פני שני ממדים – תוכן ולשון. הציון הוא סכום ציוני שני מעריכים.

מטלת כתיבה לדוגמה

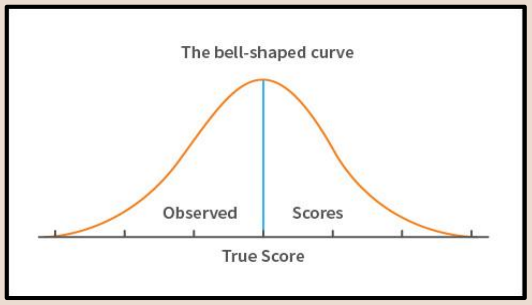
חוברת הדרכה ▪ בחינת הכניסה הפסיכומטרית לאוניברסיטאות

בעבר במרבית מדינות העולם ניתנה הזכות להצביע בבחירות לפרלמנט בגיל 21, אך זה כמה עשורים ניכרת מגמה עולמית להוריד את גיל הבחירה ל-18. בכמה מדינות, ובהן ישראל, הוגשו הצעות חוק להורדת גיל הבחירה מ-18 ל-17 ואף ל-16, ויש מדינות שכבר עשו כן. בדיון הציבורי על הורדת גיל הבחירה נשקלות שאלות כגון: מאיזה גיל אדם זכאי להשפיע על הרכב הפרלמנט, המקבל החלטות שמשפיעות השפעה ישירה על חייו? האם בני הנוער בשלים דיים להחליט החלטות המשפיעות על הציבור כולו? מה יהיו ההשלכות של הכללת בני הנוער בקבוצת הבוחרים על המעורבות של בני הנוער בנעשה במדינה ועל המערכת הפוליטית בכללה?

לדעתכם, האם יש להוריד את גיל הבחירה לכנסת? נמקו. תוכלו להיעזר בשאלות המוצגות בפסקה.

~~ נתונים ~~

- מאגר נתונים ייחודי (Cohen, 2014)
- 500 חיבורים נדגמו מקרית ממאגר גדול
- החיבורים חולקו ל-2 קבוצות של 250 חיבורים בכל אחת
- כל קבוצה נבדקה על ידי 14 ו-15 מעריכים, בהתאמה
- ממוצע ההערכות המרובות לחיבור מהווה אומדן מדויק לציונים האמתיים לכל חיבור.



~~ מדדים ~~

לכל מעריך חושבו שבעה מדדי איכות
המדדים עברו טרנספורמציה כך שהערך הגבוה יהיה הערך הטוב

מדדי דיוק

מתאם בין ציון המעריך לציון האמתי

1. ממד התוכן
2. ממד הלשון
3. סכום הציונים

פער בין ציון המעריך לציון האמתי

4. ממד התוכן (1-6)
5. ממד הלשון (1-6)
6. סכום הציונים (2-12)

ערך מוחלט

מדד יעילות (הספק)

7. מטלות ליחידת זמן (שעה)

Mean	Minimum	Maximum
.69	.41	.79
.73	.50	.83
.74	.46	.84
.29	.01	.73
.23	.00	.68
.50	.02	1.37
8.5	6.0	14.0

לא נמצא קשר מובהק בין מדדי הדיוק לבין מדד היעילות

~~ ניתוח ~~

PCA (Principal Component Analysis) ניתוח מרכיבים/גורמים ראשיים - מבטא מאגר נתונים מורכב באמצעות מספר משתנים קטן.

במחקר זה, הניתוח ביטא את מדדי האיכות – כולם או חלקם – ע"י **מדד יחיד**.

הניתוח מספק את המשקלים היחסיים של כל מדד, ומחשב את אחוז השונות המוסברת (POV) על ידי המדד היחיד, ובכך מסייע בבחירת המדדים הרלוונטיים והמודל לחישובם.

.

תוצאות 1

8 מודלים הכוללים צירוף שונה של מדדים הושוו ע"י 3 קריטריונים:

1. **POV - אחוז שונות מוסברת**

קריטריונים נוספים, מבוססים על POV ו- MVP - שונות מוסברת מצופה

2. **LIFT - העלאה** POV/MVP (**פי כמה** השונות המוסברת גבוהה מהמצופה)

3. **LEV - עוצמה** POV – MVP (**בכמה** השונות המוסברת גבוהה מהמצופה)

לא נותחו מודלים שבהם אחד המדדים מהווה סכום של שני מדדים אחרים.

הטבלה שלהלן מציגה את 2 המודלים שנבחרו

תוצאות PCA לגבי שמונה מודלים שנבדקו

Model	No. of variables	Variables	POV Proportion of variance explained	MVP Mean variance per variable	LIFT Lift	LEV Leverage
1	2	3,6	68.40	50.00	1.37	18.40
2	3	3,6,7	47.58	33.33	1.43	14.25
3	4	1,2,4,5	58.77	25.00	2.35	33.77
4	**4**	**1,2,3,6**	**75.30**	**25.00**	**3.01**	**50.30**
5	5	1,2,4,5,7	47.86	20.00	2.39	27.86
6	**5**	**1,2,3,6,7**	**61.65**	**20.00**	**3.08**	**41.65**
7	5	1,2,3,4,5	63.97	20.00	3.20	43.97
8	6	1,2,3,4,5,7	54.28	16.66	3.26	37.62

המודלים הטובים ביותר (הקריטריונים הגבוהים ביותר):

הטוב ביותר

מודל 4 – ארבעה מדדים: 3 מתאמים עם ציון אמיתי, פער בין סכום ציוני מעריך לציון אמיתי

הטוב ביותר אם מתחשבים במדד ההספק (כלומר – יש רצון לחסוך במשאבים):

מודל 6 – המדדים הנ"ל + משתנה ההספק

1. POV - אחוז שונות מוסברת
2. LIFT - העלאה – POV/MVP
3. LEV - עוצמה – POV– MVP

~~ מסקנות ~~

ניתן להגיע למדד יחיד שמבטא את המדדים השונים, עם אחוז שונות מוסברת גבוה.

תוצאות ניתוח ה-PCA מובילות למסקנות הבאות על חשיבות המדדים השונים:

1. מדדי המתאם מספקים יותר מידע ממדדי ההפרש
2. התחשבות במדד ההספק פוגעת בטיב המדד היחיד
3. משקל ההספק במודלים השונים נמוך למדי

~~ תיקוף ויישום ~~

תיקוף - התוצאות שימשו לדרוג איכותם של 24 מתוך 29 מעריכי המחקר והושוו לדירוגם על סמך נתונים דומים שחושבו מהערכות תפעוליות של מעריכים אלה. **המתאם שנמצא היה גבוה – 0.84.**

יישום - המשקלים סייעו לדרג 127 מעריכים לגביהם יש נתוני איכות דומים, כך שניתן יהיה לבצע החלטות מושכלות באשר לזימונם להערכה תפעולית, ואף לבדוק את טיב ההחלטות במחקר המשך.

המתודולוגיה שפותחה נותנת כלי לבקרת ודירוג מעריכים גם במבחני ביצוע אחרים.

ב**מחקר עתידי** כדאי לכלול מדדים נוספים (תרומה בסדנאות, זמינות, יציבות).

תודה

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

אפי ISPA ج اس
אגודה ישראלית לפסיכומטריקה
الجمعية الإسرائيلية للسيكومتريكا
ISRAELI PSYCHOMETRIC ASSOCIATION
www.ispa.org.il