

כנס אפ"י 31.1.2018
ראשון לציון

צמצום טעות המדידה של מעריכי חיבורים

יואב כהן

המרכז הארצי לבחינות ולהערכה

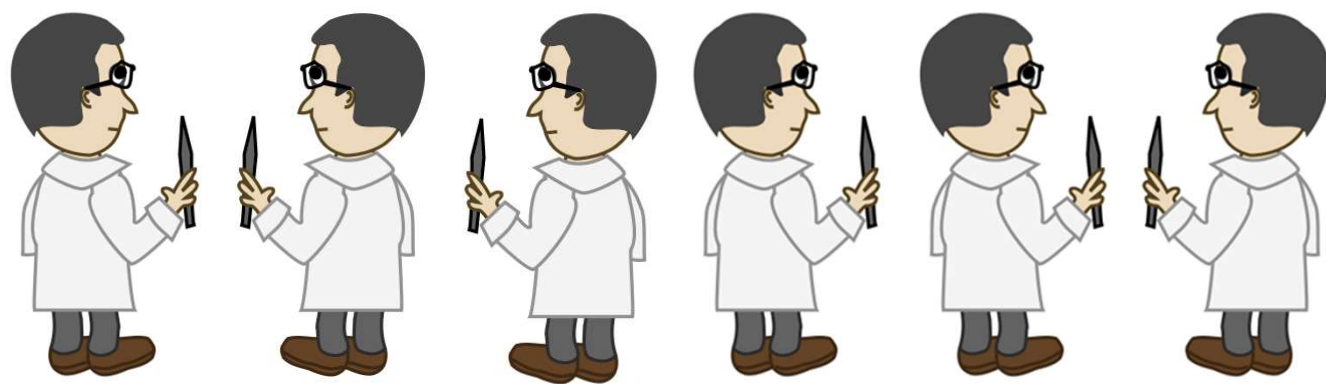
Variability of grades

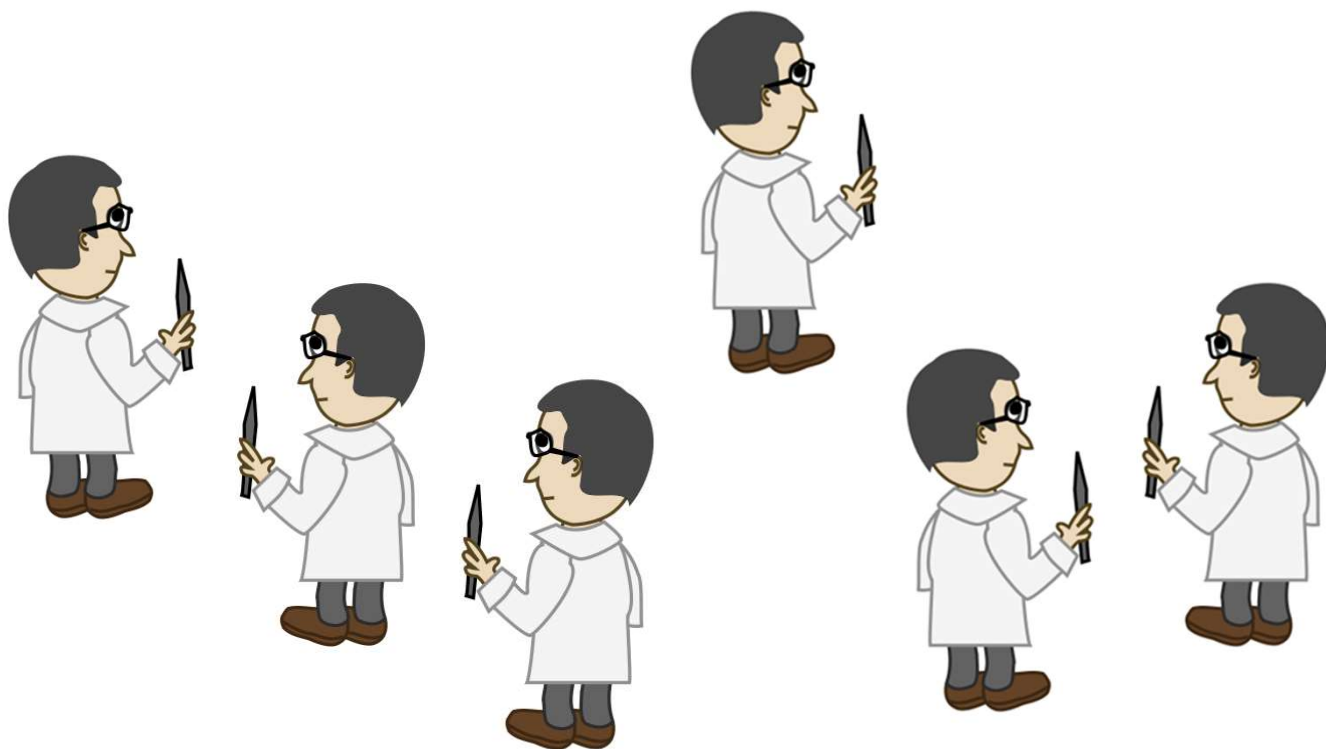
Coffman (in Ed. Measurement, 2nd edition, Thorndike 1971):

- *“different raters tend to assign different grades to the same paper” (inter-rater reliability)*
- *“a single rater tends to assign different grades to the same paper on different occasions” (intra-rater reliability)*
- *“the differences tend to increase as the essay question permits greater freedom of response”.*

מעריכי חיבורים אינם כולם שווים

- קפדנות
- שונות
- עקביות
- הטיות נוספות:
 - מגדר
 - אפקט ההילה
 - אפקט סדר
 - ...











מעריכי חיבורים אינם כולם שווים

- גם אחרי שעברו אימון ממושך
- מתבטא בציונים שהם קובעים
- עשוי להיות מקובל בהערכה כיתתית
- או בהערכה ע"י פנל
- אך לא במבחנים עתירי סיכון

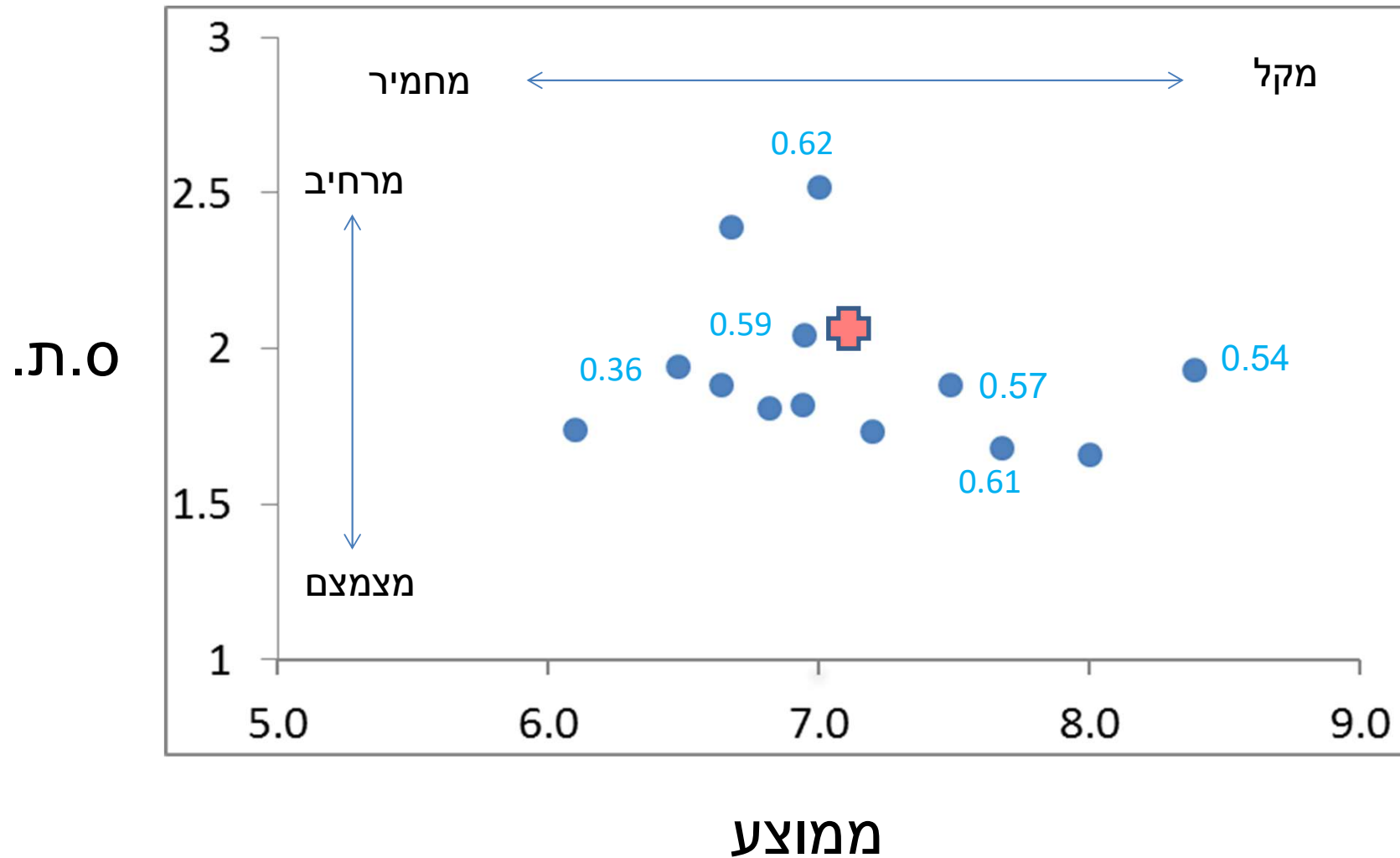
תוצאות ניסוי

- 13 מעריכים
- 250 חיבורים
- סולם ציונים 2-12

- סטטיסטיים:

- ממוצע 7.1
- סטית תקן 2.04
- מהימנות ממוצעת בין מעריכים 0.54

13 מעריכים, 250 חיבורים



Real data on 13 raters rating 250 essays

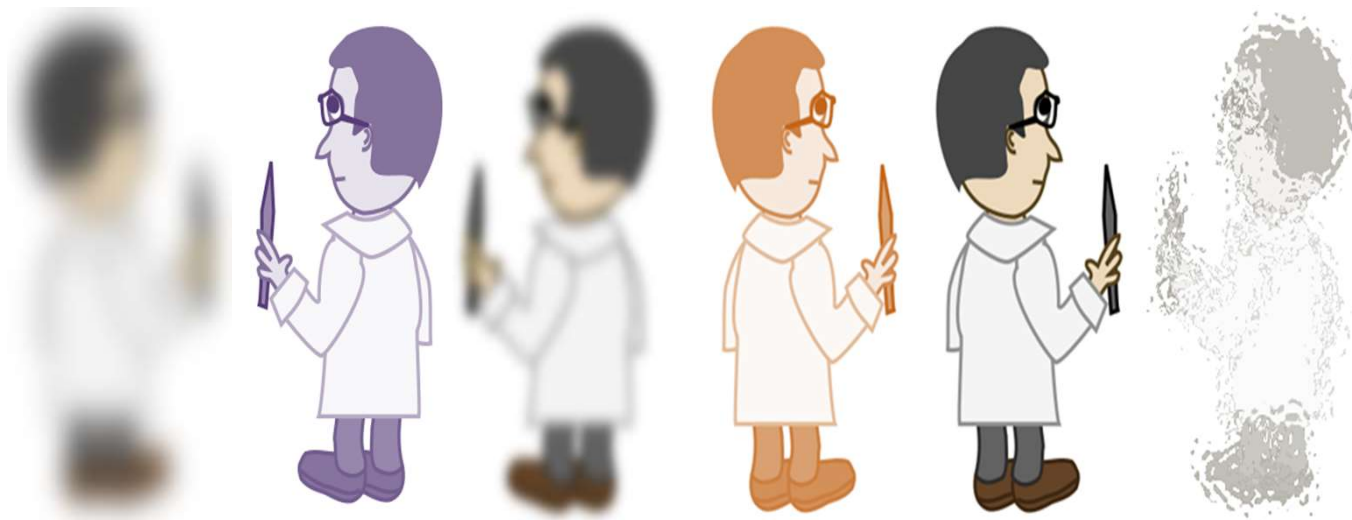
Rater	1	2	3	4	5	6	7	8	9	10	11	12	13
Mean	7.68	6.64	7.49	6.95	7.20	6.48	8.00	8.39	6.68	6.10	6.94	6.82	7.00
SD	1.68	1.88	1.88	2.04	1.73	1.94	1.66	1.93	2.39	1.74	1.82	1.81	2.52
r_{ij}	0.57	0.54	0.57	0.56	0.56	0.48	0.45	0.54	0.49	0.54	0.56	0.55	0.57

From: (Cohen & Allalouf 2016)

כיול מעריכים ← הוגנות

בהקשר של בחינות מכריעות/מבחנים עתירי סיכון,
יש לצמצם את השונות בין המעריכים על-מנת להפיק
ציונים הוגנים ומדויקים

אפשר להשיג זאת על-ידי התאמת המעריכים זה
לזה: "כיול ציונים" או: "כיול בין מעריכים"



מטרת הניסוי

• להשוות בין כמה שיטות לכיול בין מעריכי חיבורים

– ריבוי חיבורים (נבחנים) וריבוי מעריכים

– על-ידי סימולציה

– שני מעריכים לכל חיבור, על-כן שני נבחנים אינם מוערכים בהכרח על-ידי אותם מעריכים

– שיטות כיול ליניאריות

– בהנחה של שיטות הקצאה שונות של חיבורים למעריכים

על הניסוי

- שיטות לכיול ליניארי
- שיטות להקצאת חיבורים למעריכים
- תיאור הסימולציה
- תוצאות ומסקנות

כיוול ליניארי: התאמת הממוצע והפיזור (ס"ת)

הערכה מקלה



הערכה מחמירה



שיטות לכיול ליניארי

- כיול ממוצע/ס.ת.
– (“observed score linear equating”, Kolen & Brennan, 1995)
- כיול על-פי קריטריון חיצוני
– (למשל מבחן רב-ברירה)
- כיול ליניארי מרובה
– MLC - Multiple Linear Calibration (Cohen, 1993)
- בהשוואה לציונים ללא כיול

כיוול ממוצע/ס.ת. “Mean/Sigma” method

- תקנון כל המעריכים
- הנחה מובלעת: הקצאה אקראית של חיבורים למעריכים

כיוול על-פי קריטריון חיצוני

- קיים קריטריון חיצוני מהימן המקיים קשר ליניארי עם הערכות החיבורים
- לצורך הכיוול נשתמש בממוצע וסטיית התקן של הקריטריון החיצוני על אותם נבדקים המוערכים על-ידי המעריך

MLC - Multiple Linear Calibration

כיול ליניארי מרובה

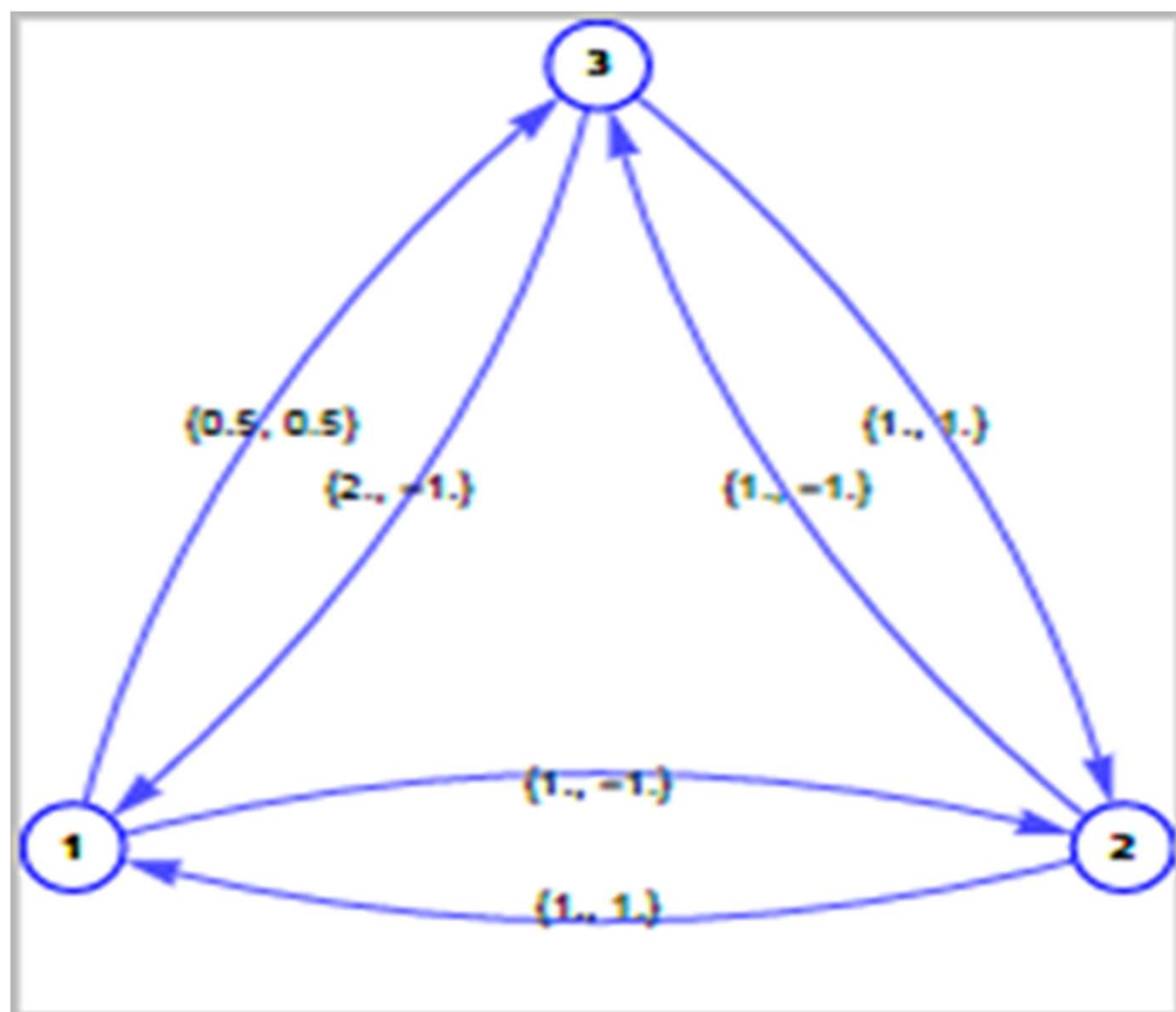
הרעיון המרכזי:

כיול כל מעריך ל"סולם גלובלי"
הסולם הגלובלי משרה כיול בין כל שני מעריכים, וכיול זה הוא הקרוב ביותר, במובן מסוים, לכיול הפשוט בין כל שני מעריכים.

יישומים:

- כיול בין מעריכים
- כיול בין ציוני קורסים
- הערכות סוציומטריות

Essay #	Rater #1	Rater #2	Rater #3
1	5	5	
2	6	4	
3	4	3	
4	4		1
5	6		3
6	5		5
7		4	3
8		5	5
9		6	4



MLC

- מצא את פונקצית הכיול המקומית f_{ij} בין i ו- j
- הנח שקיים סולם גלובלי מושלם שאליו ניתן לכייל כל מעריך. כלומר, קיימת פונקצית כיול גלובלית g_i (ופונקציה הופכית g_i^{-1})
- בהינתן פונקציות הכיול הגלובליות, ניתן למצוא פונקצית כיול גלובלית בין כל שני מעריכים: g_{ij}

MLC

כיצד מוצאים את פונקציות הכיול הגלובליות

מחפשים סט פונקציות גלובליות שעבורן הביטוי:

$$\sum_i \sum_j \sum_k d_{ijk}^2$$

מקבל ערך מינימלי.

כאשר

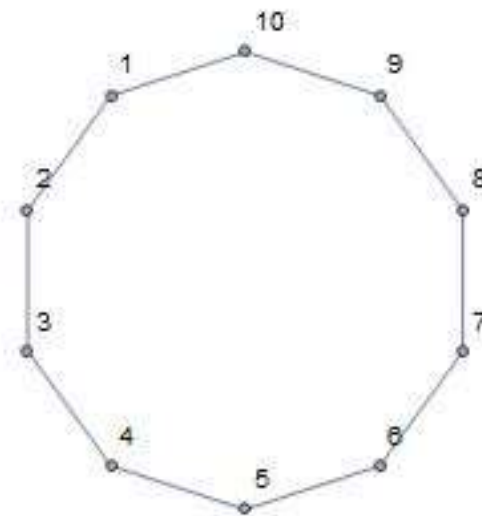
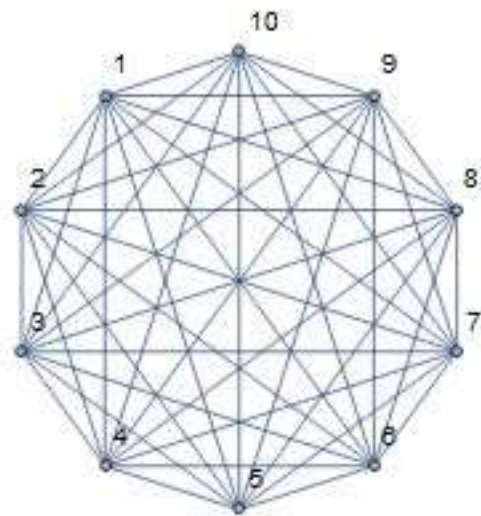
$$d_{ijk} = f_{ij}(x_{ijk}) - g_{ij}(x_{ijk})$$

עבור כל המעריכים i ו- j ועבור כל החיבורים k

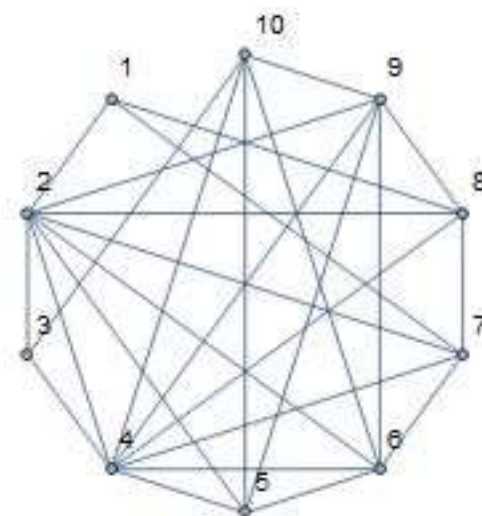
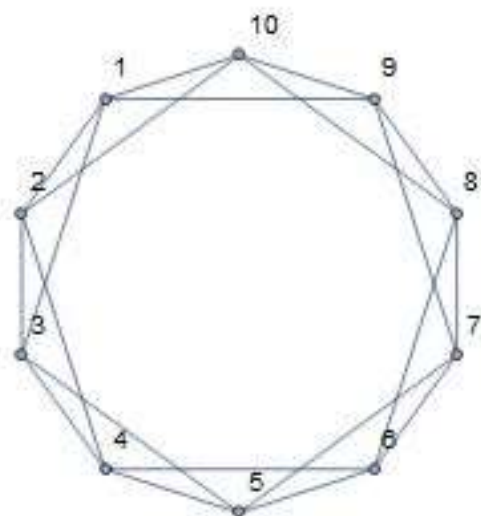
סכמות להקצאת חיבורים

כל חיבור "מוערך" "עי שני מעריכים הנבחרים ב:

- הקצאה אקראית
- הקצאת שכנים (2) (בחירת חיבורים אקראית)
- הקצאת שכנים (4) (בחירת חיבורים אקראית)
- הקצאה תוך התחשבות בציון האמיתי (בחירת חיבורים שאינה אקראית)



=



שיטה (סימולציה) 1

- **פרמטרים קבועים**

- 500 חיבורים. לכל אחד ציון אמיתי שנדגם מהתפלגות נורמלית סטנדרטית (ר' התפלגות אמפירית)
- 10 מעריכים. לכל אחד פרמטרים אישיים – ממוצע וס.ת. של הערכות (ר' טבלה)
- קריטריון חיצוני עם מהימנות של 0.9

- **20 תנאי ניסוי**

- 5 רמות של טעות מדידה (SEM) בתוך מעריך (0.4 – 0.8)
- 4 סכמות הקצאה

- **10 חזרות על כל תנאי ניסוי**

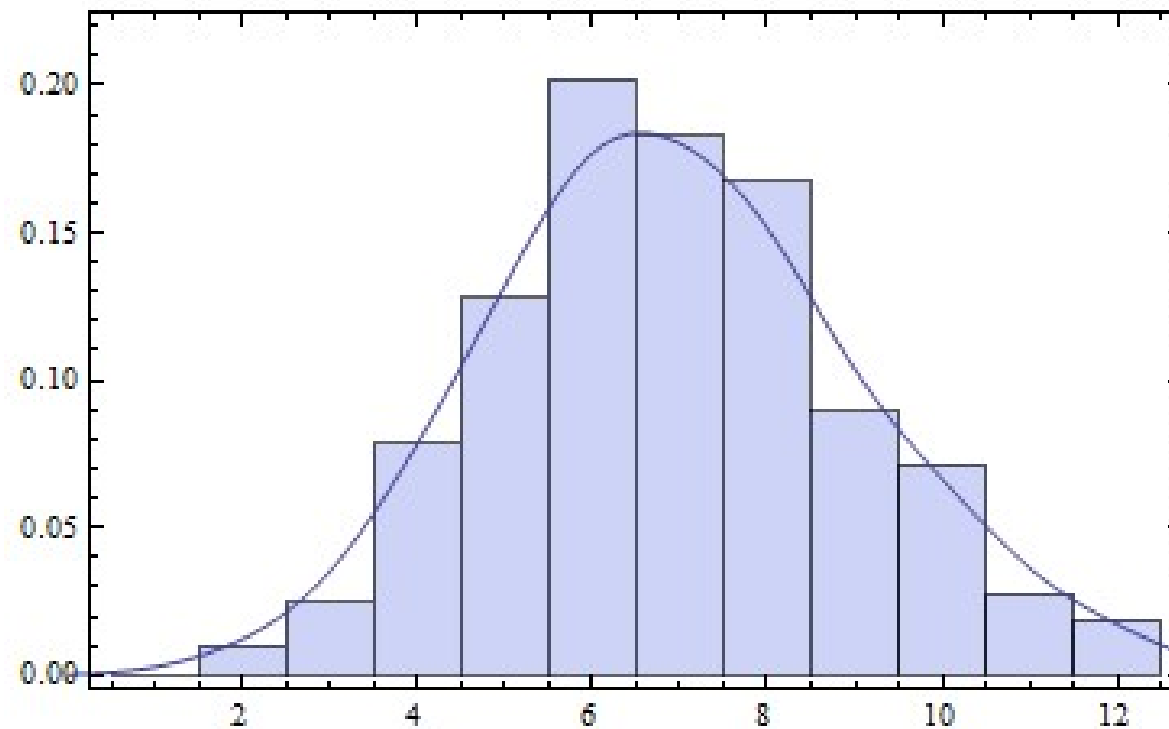
התפלגות הערכות אמפיריות

$N=7,250$

$m=6.9$

$sd=2.04$

סולם הערכה: 2-12



פרמטרים של סולמות ההערכה של 10 מעריכים

Rater #	Scale Multiplier	Scale Addend
1	1	0
2	1.2	0
3	0.8	0
4	1.2	1
5	0.8	1
6	1.2	1
7	0.8	-1
8	1.2	-1
9	1	-1
10	1	1.5

שיטה (סימולציה) 2

בכל מחזור סימולציה נאמדים 20 (18) פרמטרים
לכיוול המעריכים על-סמך כ-1000 הערכות

המשתנה התלוי:

– בכל תנאי ניסוי: טעות הכיוול

$$= \text{סכום (ציון מכויל-ציון אמיתי)}^2 = \text{MSE}$$

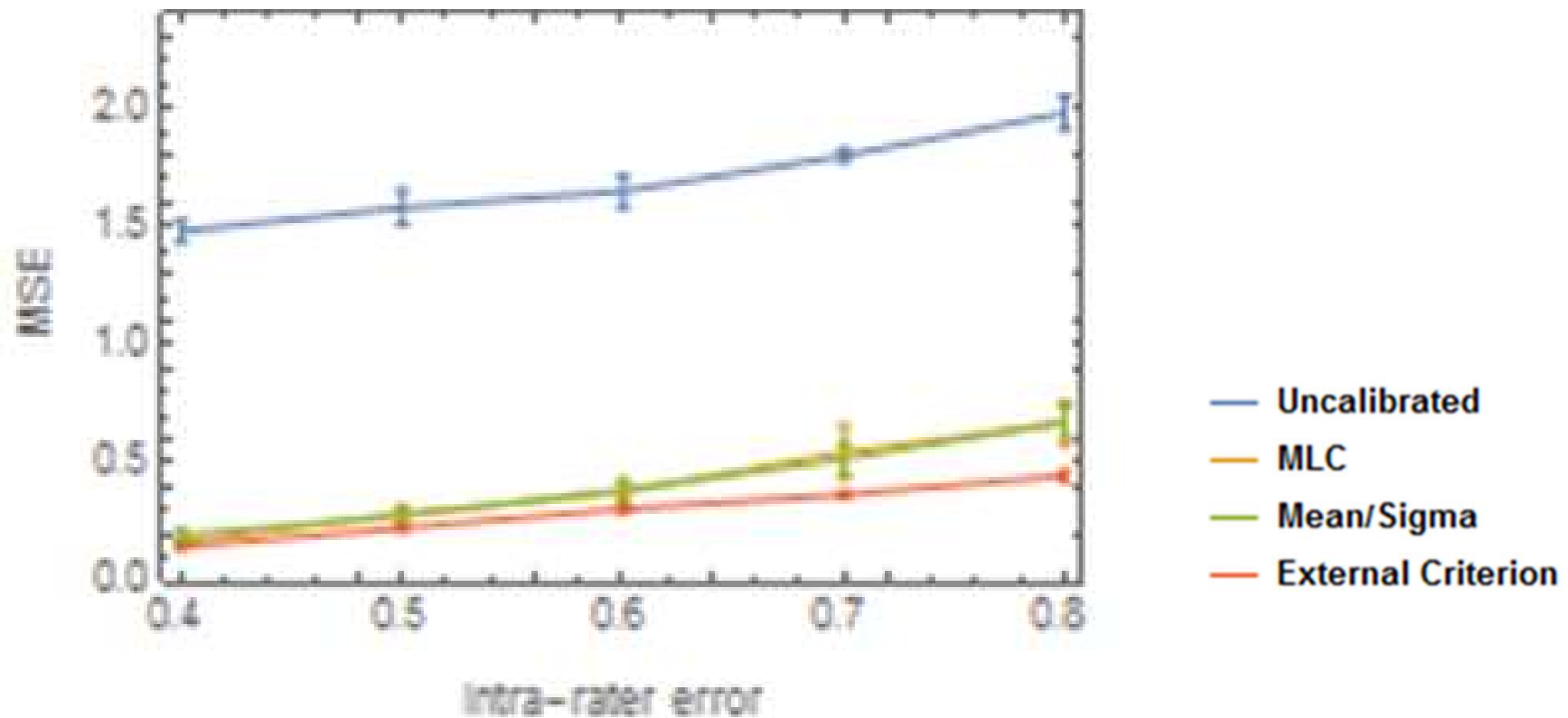
– במילים אחרות: עד כמה קרובות ההערכות המכוילות
לציונים האמיתיים

תוצאות

הטעות הריבועית הממוצעת (MSE)
כפונקציה של שיטת הכיול ומהימנות
המעריכים, בשיטות הקצאה שונות.

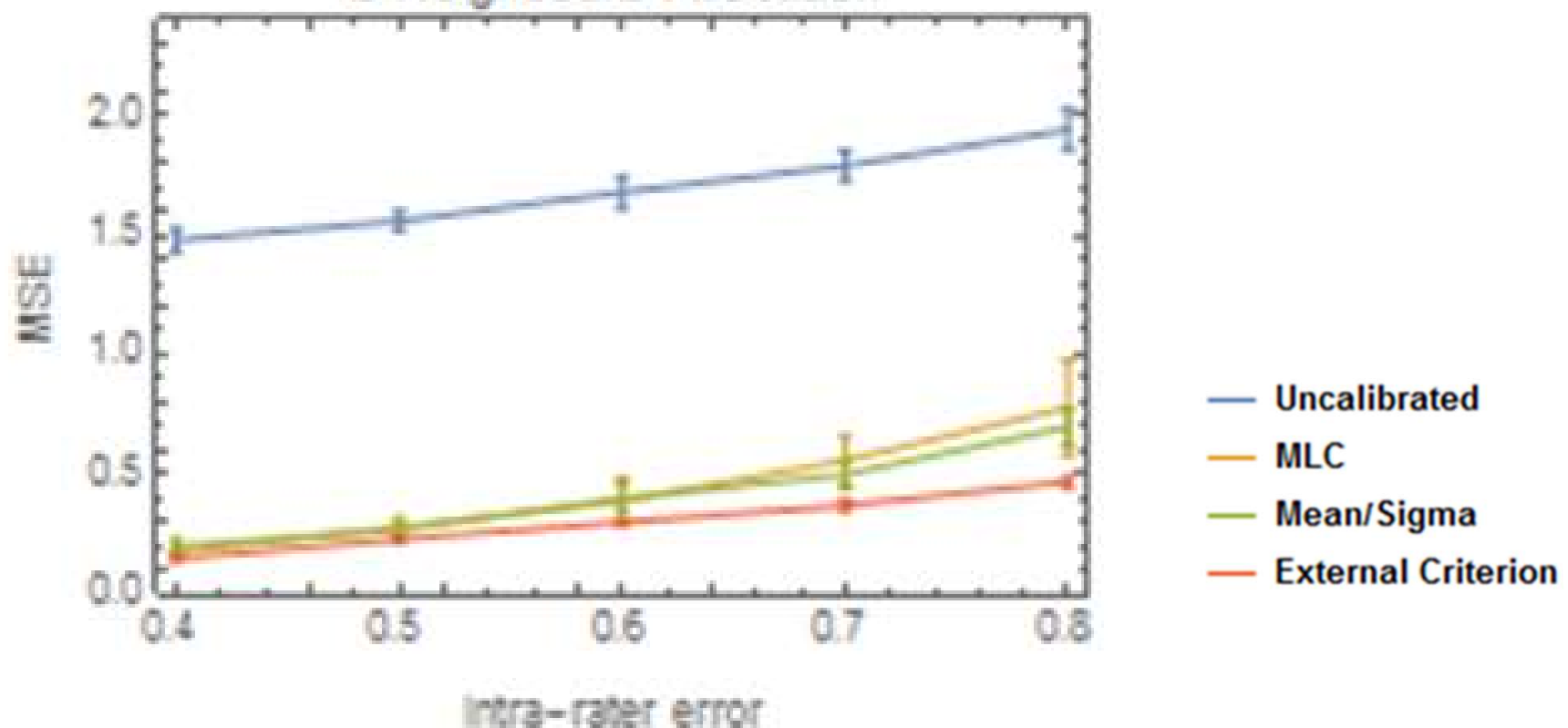
Random allocation

Calibration efficiency: 90% - 77%



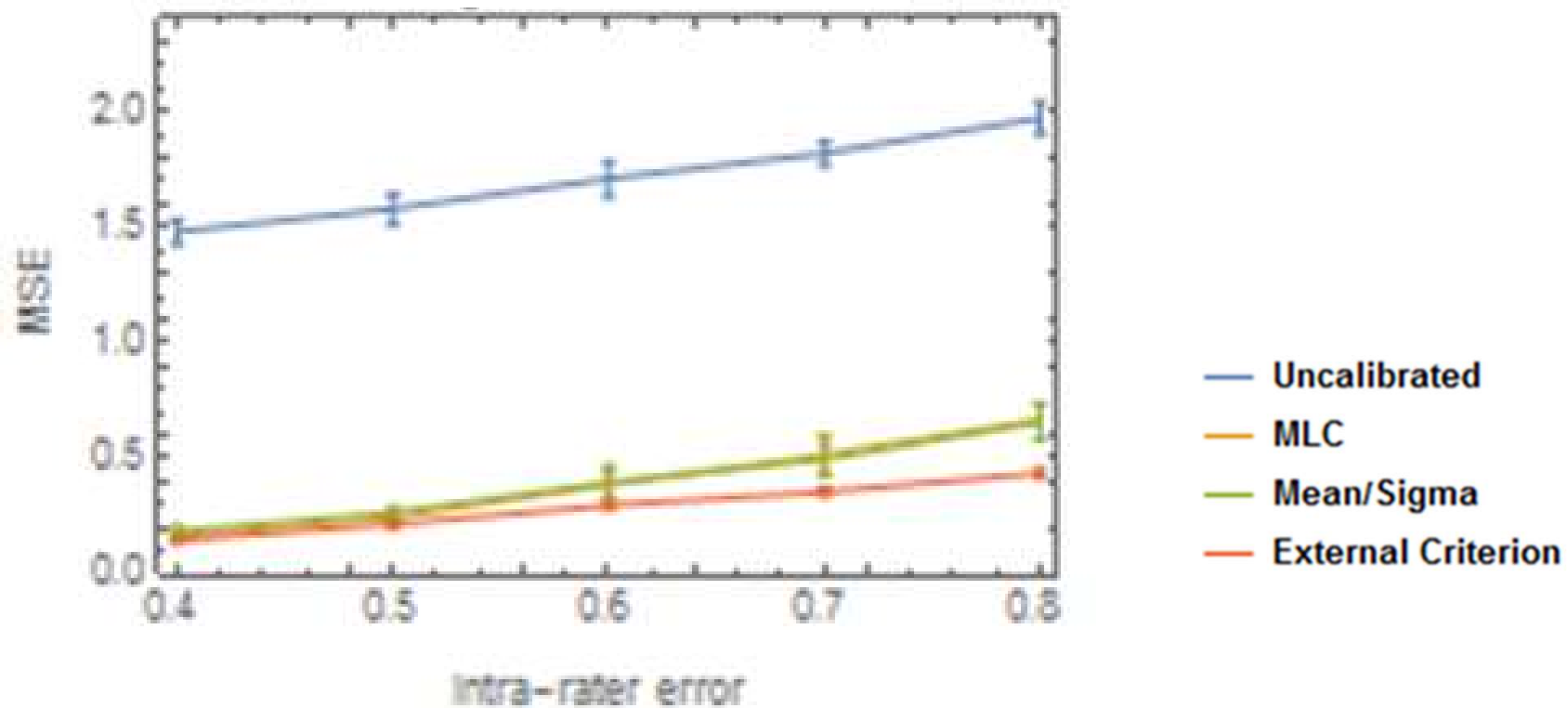
2-neighbour allocation

Calibration efficiency: 90% - 76%



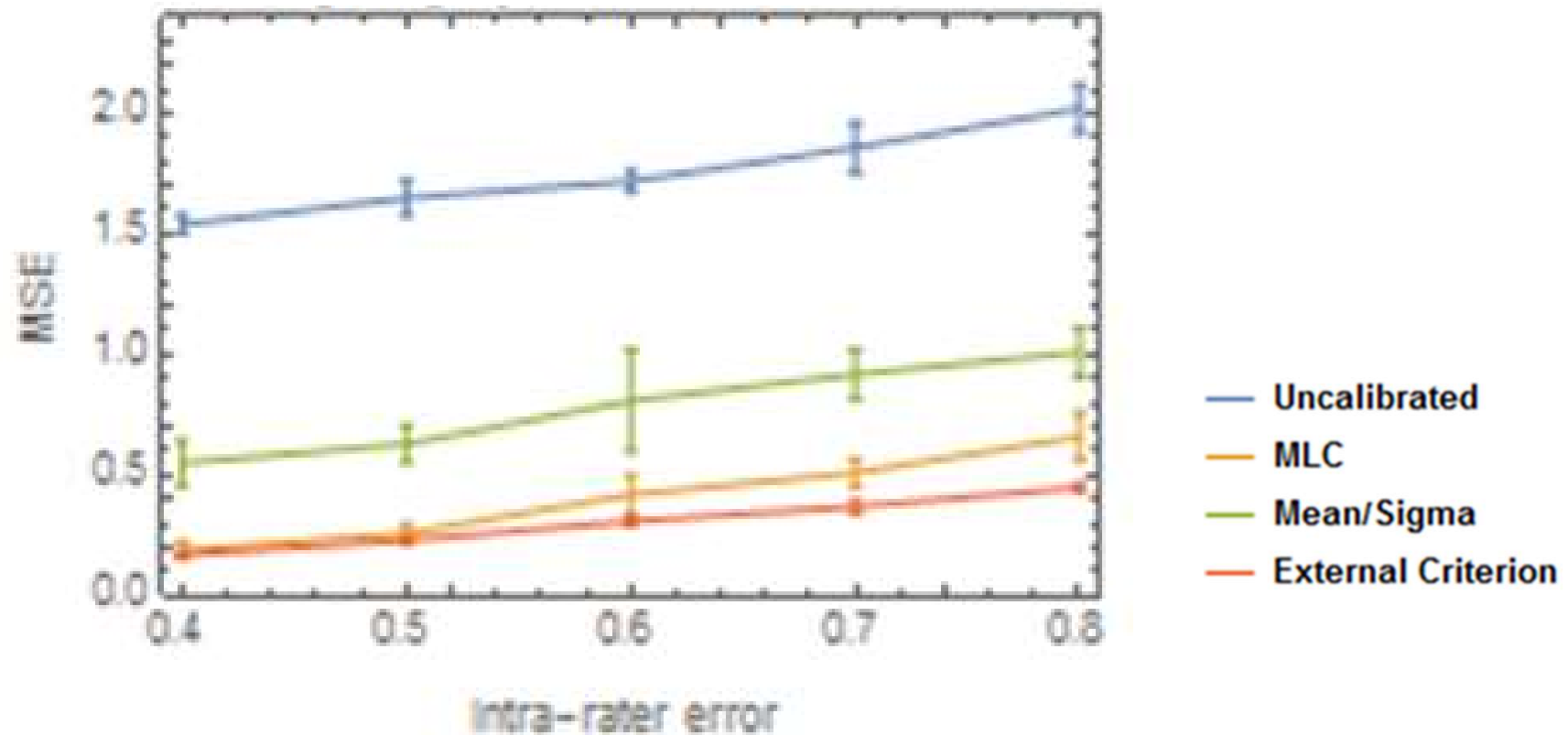
4-neighbour allocation

Calibration efficiency: 90% - 78%



Weighting by true score

Calibration efficiency: 89% - 78%



מסקנות (1)

- הערכות מכוילות **מדויקות יותר** מן הגולמיות
- הכיול יותר אפקטיבי **ברמות מהימנות גבוהות** של המעריכים(!)
- כיול על-פי **קריטריון חיצוני עדיף** על-פני כל שיטת כיול אחרת
- בהיעדר קריטריון חיצוני, עדיף להשתמש בכיול ליניארי מרובה (MLC)
- **אבל...**

האם זה תמיד נכון?

האם זה תמיד נכון שכיול על-פי קריטריון חיצוני
עדיף?

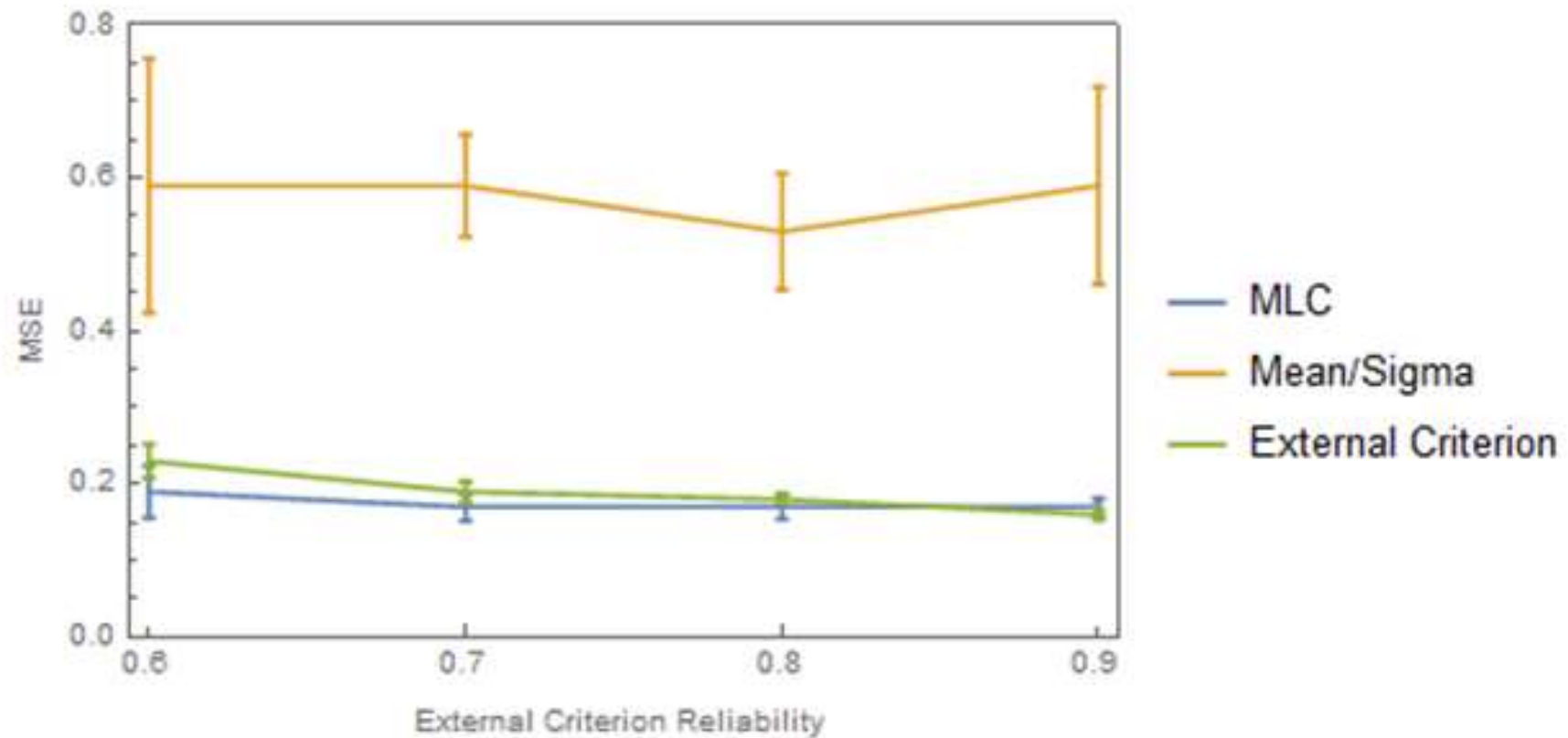
בסימולציה הנוכחית נקבעה מהימנות של 0.9
לקריטריון החיצוני

מה קורה אם הקריטריון החיצוני פחות מהימן?

בדיקה נוספת (בהקצאה שאינה אקראית)

- נועדה לוודא שהמסקנה נכונה גם כאשר ציוני הקריטריון פחות מהימנים (בהקצאה שאינה אקראית)
- מה קורה כאשר מהימנות הקריטריון החיצוני נמוכה יותר
– 0.6 או 0.7 או 0.8

אפקט המהימנות של הקריטריון החיצוני



מסקנות (2)

- הערכות מכילות **מדויקות יותר** מן הגולמיות
- הכיול יותר אפקטיבי **ככל שהמעריכים יותר מהימנים (!)**
- כיול על-פי **קריטריון חיצוני מהימן, עדיף** על-פני כל שיטת כיול אחרת
- **בהיעדר קריטריון חיצוני** (או כאשר מהימנותו נמוכה) **עדיף** להשתמש בכיול ליניארי מרובה (MLC)

ולסיכום

כיוול מעריכים



הערכה הוגנת יותר

תודה

Representing raters as a network

- We can view the inter-relations in a group of raters by representing them on a graph (in its mathematical sense).
- The vertices of the graph represent the raters, and may include any information that is unique to the rater, such as the intra-rater reliability or some other index of the quality of the rater, or statistics relating to all the essays that were rated by the rater.
- The edges between any two vertices represent the group of essays that are rated jointly by the two raters and some other information regarding the relationship between the two, such as the inter-rater correlation.
- Note that the same essay can be represented in more than one edge – when it is rated by more than two raters.
- If two particular raters do not share any essays then there is no edge connecting them.
- The graph can be a non-directed graph if the edges represent information that is the same for both raters, e.g. the number of joint essays or the correlation between the ratings of the two raters.
- It can also be drawn as a directed graph when a non-symmetric information, such as the local calibration functions (see below) between the two raters is represented by the edge.