Rate it Twice: Improving the Accuracy of Performance Evaluations Using the Wisdom of the Crowd-Within

> Meir Barneron^{1,2} Avi Allalouf¹ Ilan Yaniv²

1- The National Institute for Testing and Evaluation2- The Hebrew University of Jerusalem

The Wisdom of Crowds

The effect:

Combining estimates of multiple **different individuals** yields considerable gains in accuracy (e.g., Yaniv, 2004)

The reason:

Different individuals tend to make different errors, which cancel out when judgments are combined

The Wisdom of Crowds In the lab











The Crowd Within

The effect:

Combining multiple estimates of the **same individual** (on different occasions) yields considerable gain in accuracy (Vul & Pashler, 2008; Herzog & Hertwig, 2014)

The reason:

Judgments made on different occasions "within the same person" involve different sources of errors that cancel out each other when combined

The Crowd Within In the lab











Main Goal

Investigate the crowd-within in the human

performance evaluations.

We concentrate on the essay evaluations in the

psychometric entrance test.

כי ברוץ הארושי היא המינה לו הארוני לעניה היא לו כין דרך הקול היא לה של הארון להיין לארושי היון לארושי היין להיי לארשי היין להיי לארשי היין להיין לארשי היין להיי לארשי היין להיי לארשי היין להיין לארשי היין ל בארשי היין לארשי היין לארשי היים לארשי לארשי לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין ל בארשי היין לארשי היין לארשי היים לארשי לארשי לארשי לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין לארשי בארשי היין לארשי היין לארשי היים לארשי לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין לארשי בארשי היין היין לארשי היין לארשי הארשיה אוויין לארשי היין לארשי היין לארשי היין לארשי היין לארשי היין לארשי היי לארשי היין לארשי היין לארשי הארשיה אווין לארשי הארשיה אווין לארשי היין לארשי לארשי היין לארשי היין לארשי היין לארשי היין לארשייין לארשי היין לארשיי היין לארשי היין לארשי היין לארשי היין לארשייין לארשייין לארשי היין לארשי היין לארשי היין לארשייין היינייין לארשייין לארשיי לארשייין לארשייין לארשיייין לאר

ي مرتبع مرتبع مرتبع المالي من مرتبع المريس من محمد بنا الله معط مريس مرتبع مرتبع مرتبع المرابع مرتبع المريس مرتبه مالك المريس مرتبع مرتب مرتب مرتب مرتب مرتب مرتبع مرتبع مرتبا مرتبي مرتبع مرتبع مرت

"אמרים בעניי בן באך בירה לידוד את על בבריד לענים, אך אפר אלי הדעי, האריקה את לאי הלא לעניי לי שיר ער עיידלעים העניים אין על אל נדע אין אין "לארי ליפר לעניי און אילי היי הארידים... על אל בעניים היים לא מרכים בעני הארי בעני הדעי עניי הארידים בעניים לעל בעניים היים לא מרכים בעניים אינים על היים או בעניים או היים בארידים. בנוסקראה עין קליום לעניים אל אין.





The Essay Task

Since 2012, candidates are required to write a short essay (25-50 lines)

Essays are rated on two 6-point-scales (1- poor, 6excellent)

- Content (thesis development; coherency; critical thinking)
- Language (fluency; precise use of the words, grammar and syntax; sentence structure complexity; use of linguistic tools to organize the text)
- Final score = content + language (2- poor, 12excellent)

The Essay Task

Common method:

Two **different raters** evaluate each essay. The two grades are then combined.

Tested method:

The **same rater** evaluates each essay twice. The two grades are then combined.

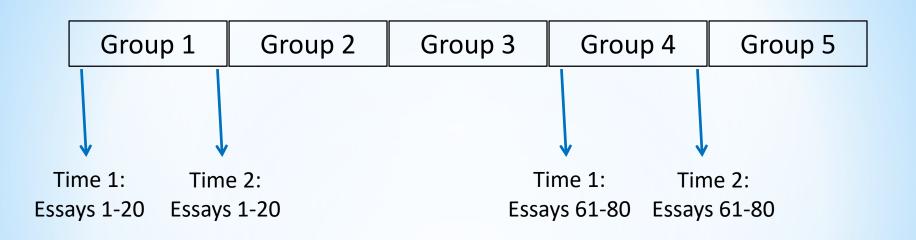
Method Participants and procedure

Participants 100 essays, 30 raters

<u>Procedure</u>

- I. 3-hour workshop
- II. Each rater evaluated 20 essays (Time 1)
- III. Re-evaluated the same essays 1 week later (Time 2)

Method Design



<u>Withbreratfeevalundtioantsio29</u> (raters)* 20 (essays per session) * 2 (Assessing ass) f that gis ald @ 60 t Time 1 and Time 2

<u>Different-rater combination</u> Average of the grades of two randomly selected raters at Time 1



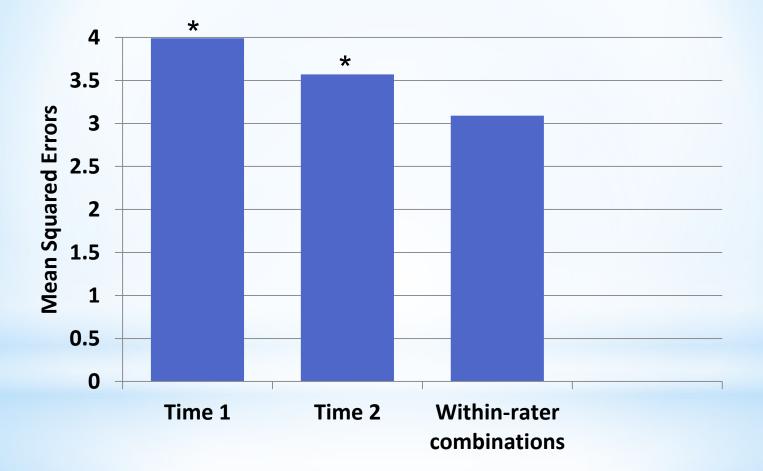
Empirical True Scores

Average of the grades of 15 raters who rated the same 100 essays in the past (Cohen, 2015).

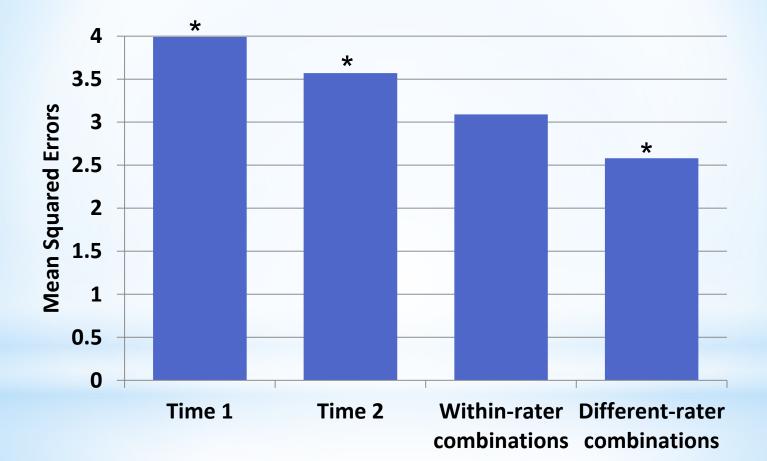
Measures of Accuracy

- <u>Squared Errors</u> (per rater, essay)
 Squared distance between the "criterion" and the grade, at Time 1, Time 2.
 Within-rater combinations should yield **lower** squared errors than both the grades at Time 1 and Time 2.
- II. <u>Correlation with True Scores</u> (per rater) Correlation between the "true score" and the evaluation, at Time 1, Time 2. Within-rater combinations should yield **higher** correlations with the criterion than both the grades at Time 1 and Time 2.

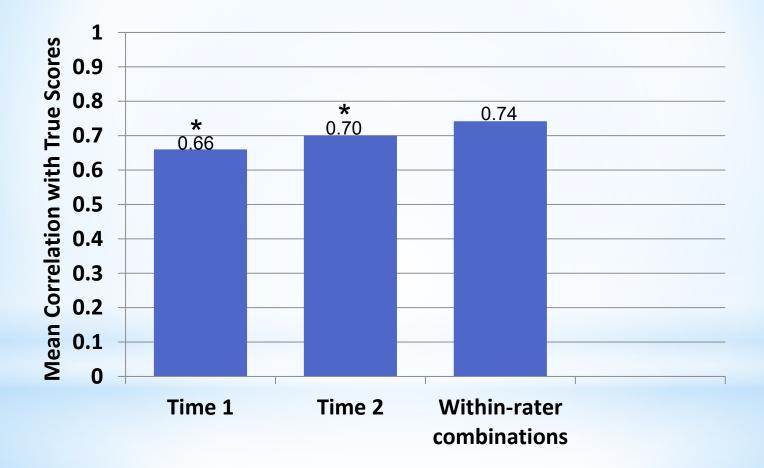




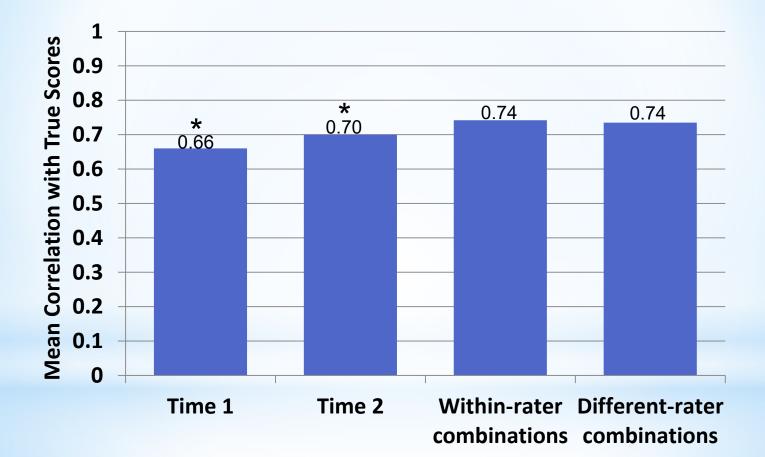




Results-Correlations with True Scores



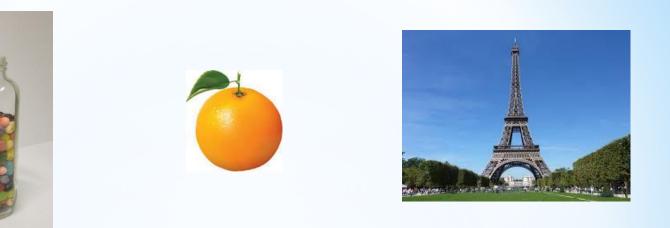
Results-Correlations with True Scores





- Within-rater combinations improve accuracy in terms of both squared distance from the true score and correlation with true score
- Different-rater combinations perform even better (in term of squared errors)

Theoretical contributions







به موجود میروند. میروند کافت یکی چار عدما کنونی بوراند از ۲۰۰۰ میرونی امراضی کار هوری و به امور میرونانوند، بوراند امور حال ایر روین میروند امام کار میروند امور ایرون میروند. میرون میروند امام معرفین میروند امام معرفین میروند میروند میروند. ۲۰۰۰ مورد میروند امور با ایرون میروند میروند امام میرون میروند اورونا

- In a new domain (performance evaluations)
- When the criterion is (also) subjective
- With complex stimulus
- Using experts

Practical implications

Other domains:

 performance evaluations (e.g., students evaluations, job interviews, resume etc.)

Systematic use of within-judge combinations could

- Improving evaluations accuracy
- Yield financial savings

Limitations and future directions

We assume that the two evaluations are (somewhat) independent.

Is it always the case?

(Few evaluations, short time between evaluations etc.)

Concluding Remark

"תחשוב שנית"

"Think twice"

"Il faut tourner sa langue 7 fois dans sa bouche avant de parler"

"7 раз отмерь один раз отрежь"

Thanks to...

NITE Research Fund

Avi Allalouf Ilan Yaniv Jeny Shmulevich

Thank You!

E-mail: meir@nite.org.il

