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

Meir Barneron[1,2]
Avi Allalouf[1]
Ilan Yaniv[2]

1- The National Institute for Testing and Evaluation
2- 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.

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, 6- excellent)

- 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, 12- excellent)

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

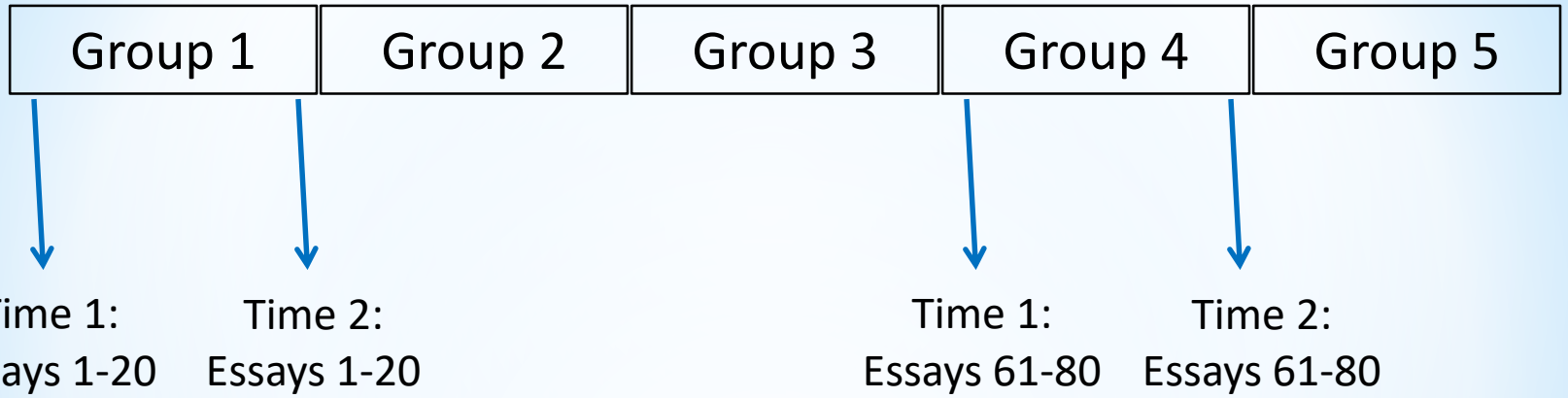
100 essays, 30 raters

Procedure

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



With-rater combination:

Number of evaluations: 29 (raters)* 20 (essays per session) * 2 (sessions) = 1,160

Average of the grades at Time 1 and Time 2

Different-rater combination

Average of the grades of two randomly selected raters at Time 1

Criterion

Empirical True Scores

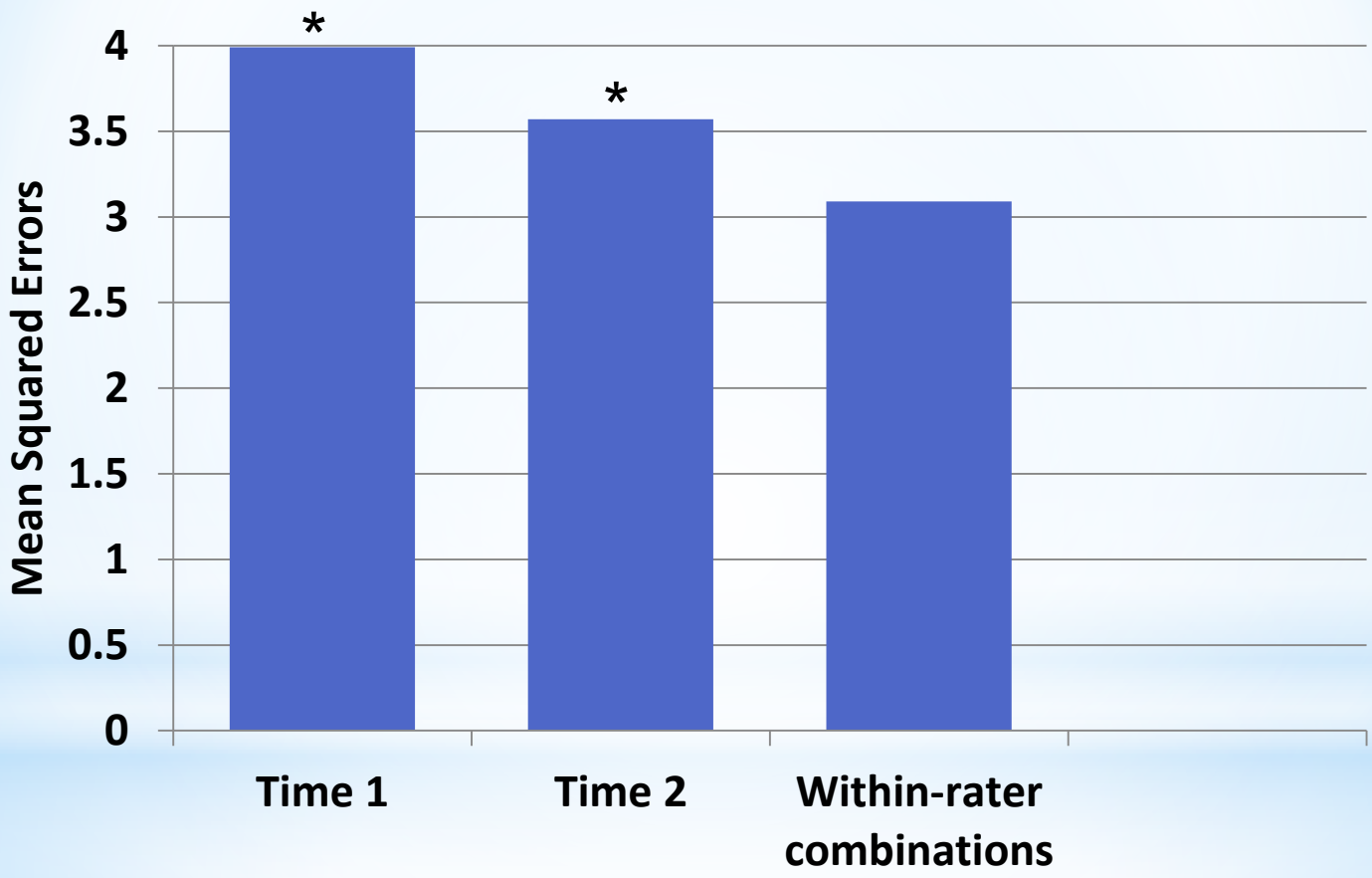
Average of the grades of 15 raters who rated the same 100 essays in the past (Cohen, 2015).

Measures of Accuracy

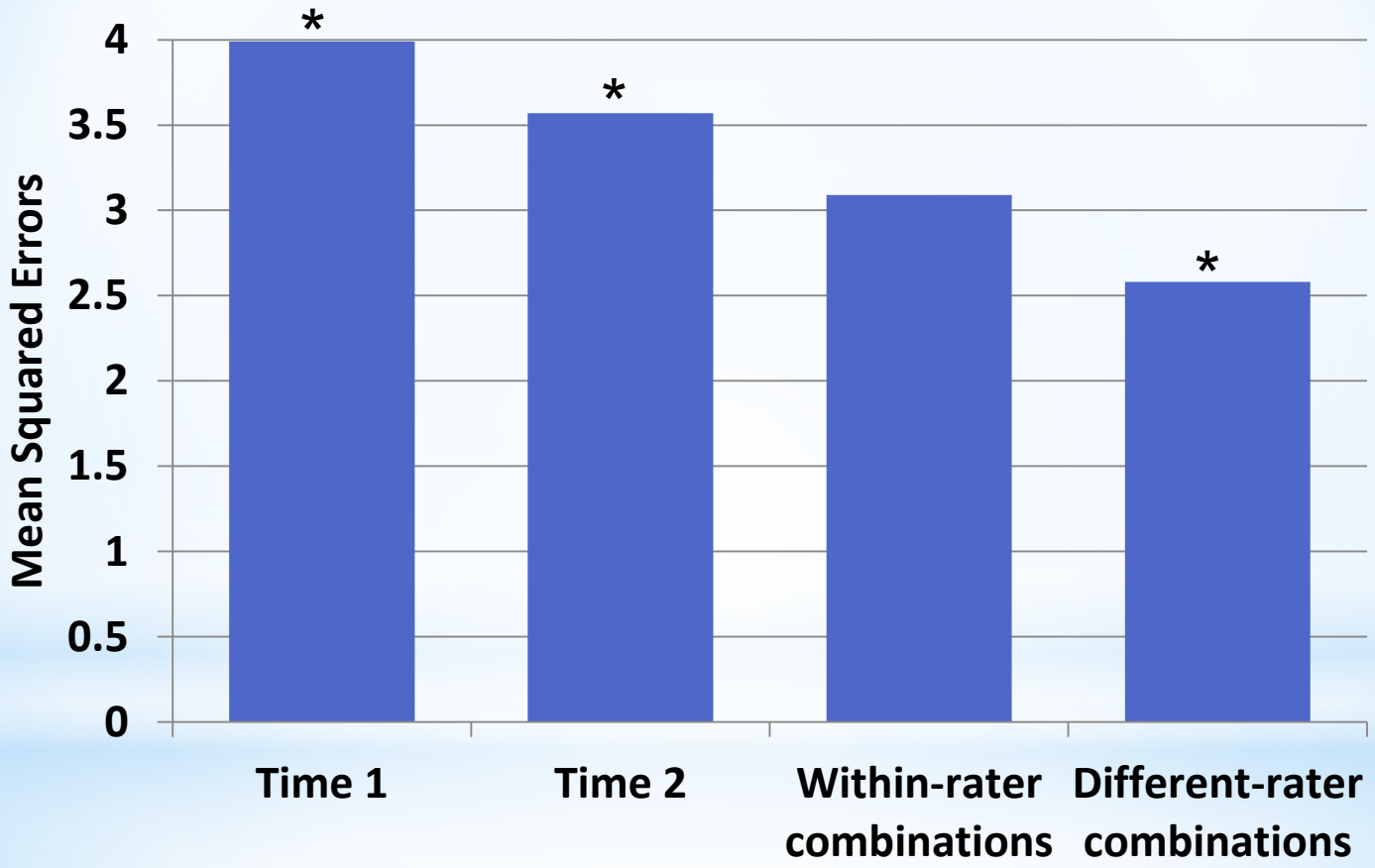
I. Squared Errors (per rater, essay)
Squared distance between the "criterion" and the grade, at Time 1, Time 2.
Within-rater combinations should yield **lower** squared errors than both the grades at Time 1 and Time 2.

II. Correlation with True Scores (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.

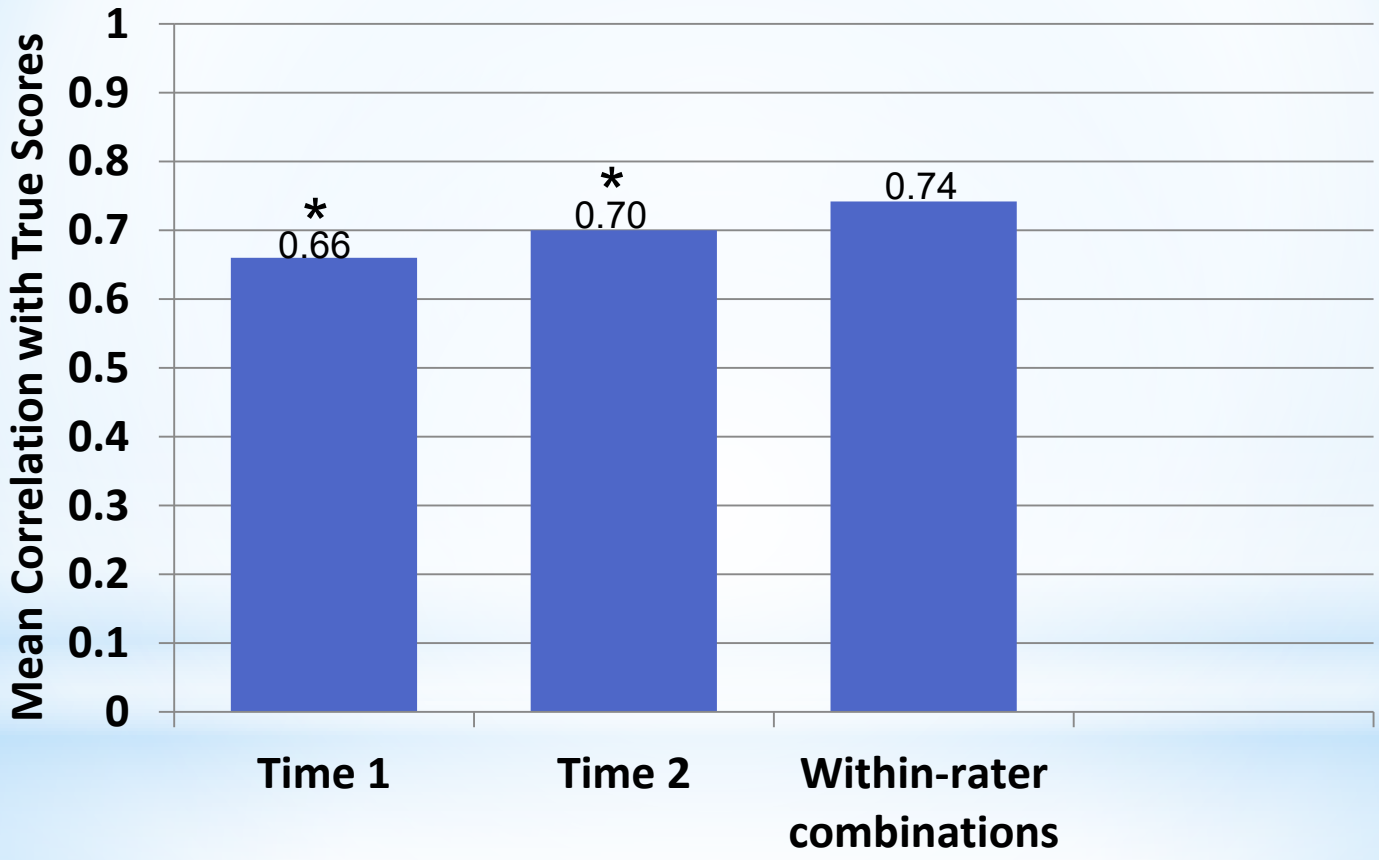
Results-
Mean Squared Errors



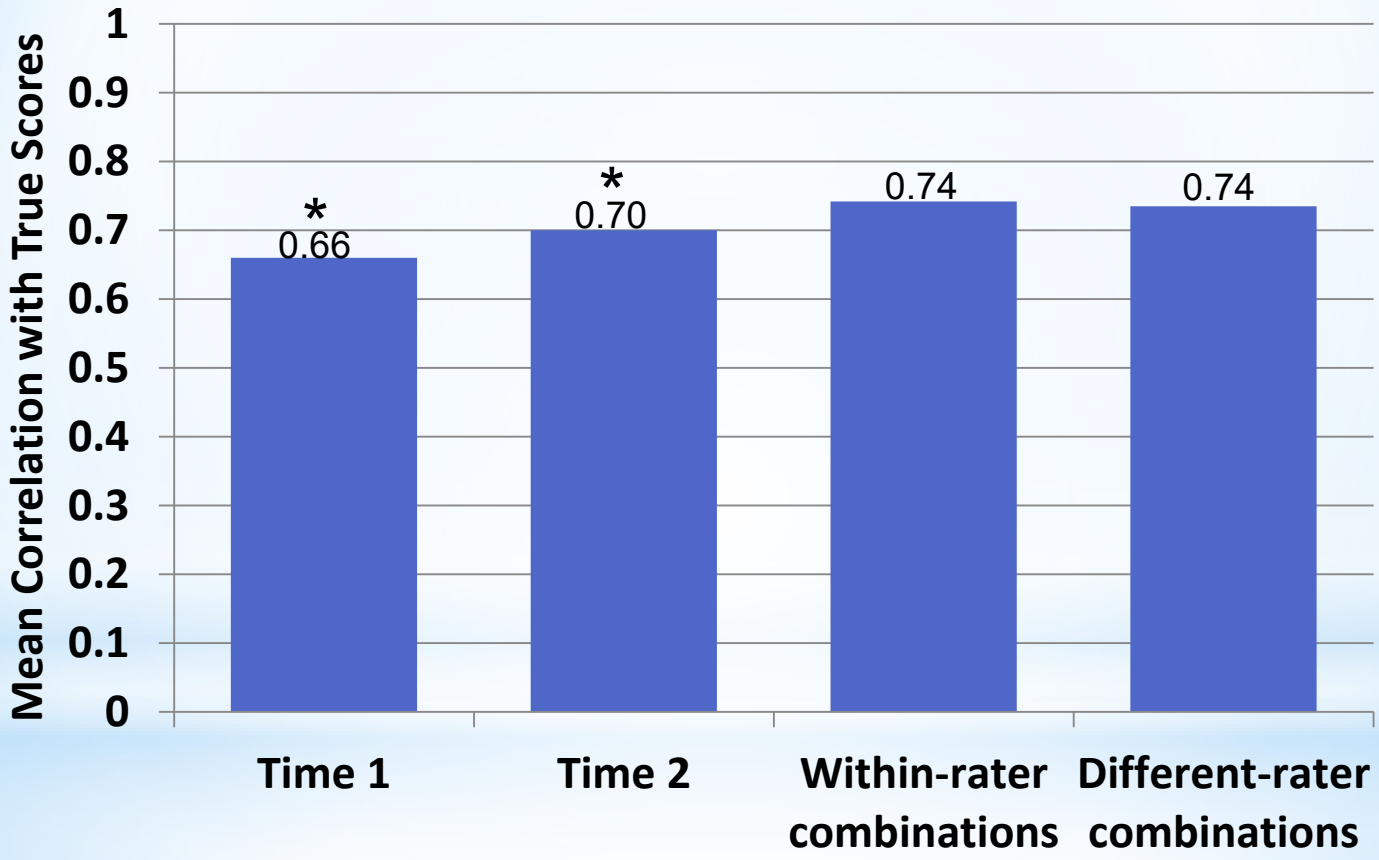
Results-
Mean Squared Errors



Results-
Correlations with True Scores



Results-
Correlations with True Scores



Summary

- 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

Testing the crowd-within

- 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