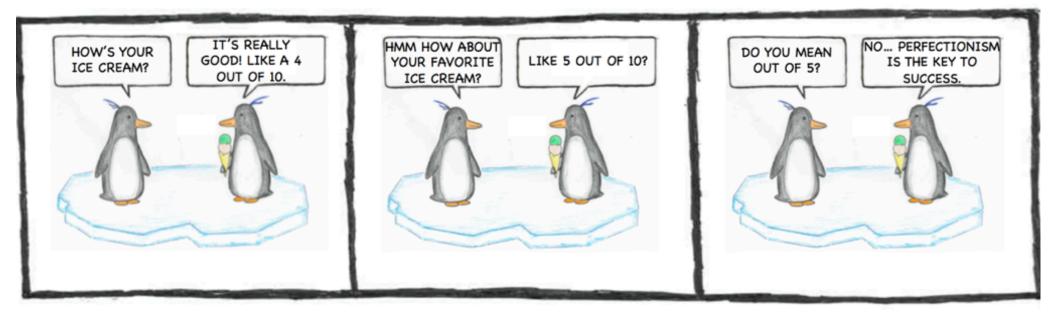
# Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings

Jingyan Wang, Nihar B. Shah Carnegie Mellon University



## Miscalibration

### People have <u>different scales</u>

### when giving <u>numerical scores</u>.

Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings

Jingyan Wang Camegie Melion Univer Jingyarwijkos.cmu.ed

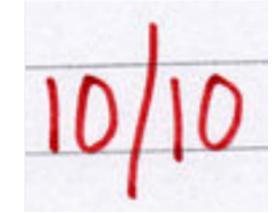
ABSTRACT A key eap in building multi-agent systems in to gather data suoreducid by the agents (propic), in either cardinal (sumeric natings) or endual (pashings) from Cardinal enerse onlented from propite will have to tradiffe from misoidistics. A propher approach to address the issues in to assues any data: model of macalitetation to a linear bisers of the bisers from Sing Approach. Neverex.



Nihar B. Shah Camegie Mellon Univer niharsijtes ems-edu

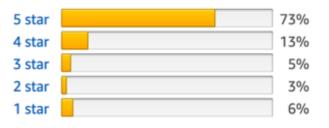
review is sugged on distant time for evaluation, and this argument is down sufficient by the random. There we reviewes providing evaluations then the start  $2\times 3$  for the respective for the start  $2\times 3$  for the respective  $3\times 3$  for the re

that of minishlawinis – the systematic errors introduced due to comparability of contained scores provided by different propile rule [3, 3], and references theories]. This issue of minishlawini is insertient minishlawing and gaugglying gaugangiane. about the form of manifellawing, and due comparements per reviewer addition biases [2, 16, 20]. The system of the score asymptotic score (3, 16), and the one parameteris per reviewer addition biases [2, 16, 20]. The propile of the score asymptotic score (3, 16), and the propile of the score asymptotic score (3, 16), and the propile of the score asymptotic score (3, 16), and we due to application of the score asymptotic score (3, 16), and we due to application of the score asymptotic score (3, 16), and we due to application of the score (3, 16) and the score (3, 16).



#### 4,072 customer reviews

🖢 🚖 🊖 🥎 🛛 4.3 out of 5 stars ~

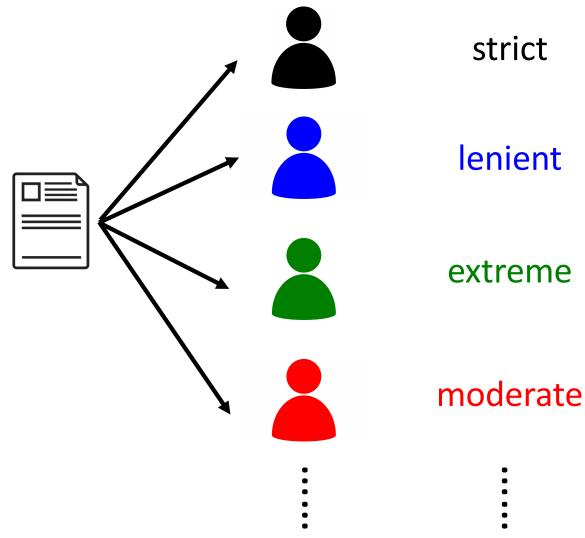


#### reviewing papers

#### grading essays

rating products

## People are miscalibrated



Wang & Shah

#### **Arbitrary Miscalibrations in Ratings**

## Miscalibration

#### • Ammar et al. 2012

"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another."

### • Mitliagkas et al. 2011

"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."



What should we do with these scores?

# Two approaches in the literature

### 1. Assume simplified models for calibration

[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba and Kashima 2013, Ge et al. 2013, MacKay et al. 2017]

- People are complex [e.g. Griffin and Brenner 2008]
- Did not work well in practice: *"We experimented with reviewer normalization and generally found it significantly harmful."* — John Langford (ICML 2012 program co-chair)

### 2. Use rankings

[Rokeach 1968, Freund et al. 2003, Harzing et al. 2009, Mitliagkas et al. 2011, Ammar et al. 2012, Negahban et al. 2012]

- Use rankings induced from the scores or directly collect rankings
- Commonly believed to be the only useful information, if no assumptions on calibration

## Folklore belief

#### Freund et al. 2003

"[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences."

?; ?

Is it possible to do better than rankings with essentially no assumptions on the calibration?

# Simplified setting



# $x_A \in [0, 1]$

 $x_{R} \in [0, 1]$ 



Calibration function  $f_1: [0, 1] \rightarrow [0, 1]$ Gives score  $f_1(x_i)$  for  $i \in \{A, B\}$ 



Calibration function  $f_2: [0, 1] \rightarrow [0, 1]$ Gives score  $f_2(x_i)$  for  $i \in \{A, B\}$ 

- $f_1, f_2$  are strictly monotonic
- Adversary chooses  $x_A$ ,  $x_B$  and strictly monotonic  $f_1$ ,  $f_2$
- Papers assigned to reviewers at random

# Simplified setting

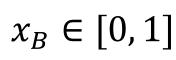


## $x_A \in [0, 1]$



Calibration function  $f_1: [0, 1] \rightarrow [0, 1]$ Gives score  $f_1(x_i)$  for  $i \in \{A, B\}$ 







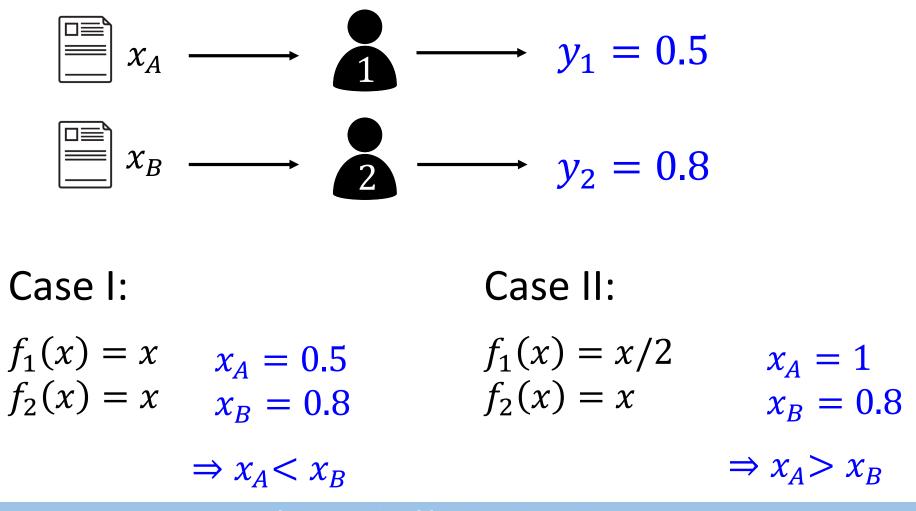
Calibration function  $f_2: [0, 1] \rightarrow [0, 1]$ Gives score  $f_2(x_i)$  for  $i \in \{A, B\}$ 

- Goal: infer  $x_A > x_B$  or  $x_A < x_B$ ?
- Eliciting ranking vacuous: random guessing baseline
- $y_i$  denotes score given by reviewer  $i \in \{1, 2\}$

Given  $\{y_1, y_2, assignment\}$ , is it possible to infer  $x_A > x_B$  or  $x_A < x_B$  better than random guessing?

## Impossibility?

Intuition: The reported scores can be either due to x, or due to f.



**Arbitrary Miscalibrations in Ratings** 

## Impossibility... for deterministic algorithms

**Theorem:** No **deterministic** algorithm can always be strictly better than random guessing.

• Stein's paradox

[Stein 1956]

• Empirical Bayes

[Robbins 1956]

• Two envelope problem [Cover 1987]



## Proposed algorithm

Algorithm: The paper with the higher score is better, with probability  $\frac{1+|y_1-y_2|}{2}$ .

**Theorem:** This algorithm uniformly and strictly outperforms random guessing.

#### Scores > rankings!

Algorithm: The paper with the higher score is better, with probability  $\frac{1+|y_1-y_2|}{2}$ .



Algorithm: The paper with the higher score is better, with probability  $\frac{1+|y_1-y_2|}{2}$ .

	$f_1$	
$x_A = 0$	0.1	
$x_B = 1$	0.3	

Algorithm: The paper with the higher score is better, with probability  $\frac{1+|y_1-y_2|}{2}$ .

	$f_1$	$f_2$
$x_A = 0$	0.1	0.5
$x_B = 1$	0.3	0.9

Wang & Shah

Algorithm: The paper with the higher score is better, with probability  $\frac{1+|y_1-y_2|}{2}$ .

	$f_1$	$f_2$
$x_A = 0$	0.1	0.5
$x_B = 1$	0.3	0.9

- Under blue assignment, output paper B with probability  $\frac{1 + |0.1 0.9|}{2} = 0.9$
- Under red assignment, output paper A with probability  $\frac{1 + |0.3 0.5|}{1 + |0.3 0.5|} = 0.6$

• On average, correct with probability  $\frac{0.9 + (1 - 0.6)}{2} = 0.65 > 0.5$ 

### Extensions

- A/B testing and ranking
- Noisy setting

## Take-aways



### • Scores > rankings

in presence of arbitrary miscalibration



### Randomized decisions

good for both inference and fairness

[Saxena et al. 2018]

