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 **different scales** when giving **numerical scores**.

- reviewing papers
- grading essays
- rating products

**4,072 customer reviews**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>3,016</td>
<td>73%</td>
</tr>
<tr>
<td>4 star</td>
<td>544</td>
<td>13%</td>
</tr>
<tr>
<td>3 star</td>
<td>203</td>
<td>5%</td>
</tr>
<tr>
<td>2 star</td>
<td>127</td>
<td>3%</td>
</tr>
<tr>
<td>1 star</td>
<td>106</td>
<td>6%</td>
</tr>
</tbody>
</table>

- 4.3 out of 5 stars
People are miscalibrated

- strict
- lenient
- extreme
- moderate

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
   - 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
   - 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

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

$x_B \in [0, 1]$  
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

- 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 $\{y_1, y_2, \text{assignment}\}$, is it possible to infer $x_A > x_B$ or $x_A < x_B$ better than random guessing?
Intuition: The reported scores can be either due to x, or due to f.

Case I:
\[ f_1(x) = x \quad x_A = 0.5 \quad x_B = 0.8 \]
\[ \Rightarrow x_A < x_B \]

Case II:
\[ f_1(x) = x/2 \quad x_A = 1 \quad x_B = 0.8 \]
\[ \Rightarrow x_A > x_B \]
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!
Intuition

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

<table>
<thead>
<tr>
<th></th>
<th>( x_A = 0 )</th>
<th>( x_B = 1 )</th>
</tr>
</thead>
</table>

Intuition

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

<table>
<thead>
<tr>
<th>$x_A = 0$</th>
<th>$f_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_B = 1$</th>
<th>$f_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>
Intuition

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

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_A = 0$</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>$x_B = 1$</td>
<td>0.3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Wang & Shah

Arbitrary Miscalibrations in Ratings
Intuition

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

<table>
<thead>
<tr>
<th>( x_A = 0 )</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_B = 1 )</td>
<td>0.3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

• **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
  \[
  \frac{1 + |0.3 - 0.5|}{2} = 0.6
  \]

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

**Wang & Shah**

*Arbitrary Miscalibrations in Ratings*
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]
Thanks! Questions?