Improving Peer Ratings Using Pairwise Comparisons and Elo Scoring

Ian Kloo
Data Scientist/Assistant Professor
Department of Systems Engineering
United States Military Academy

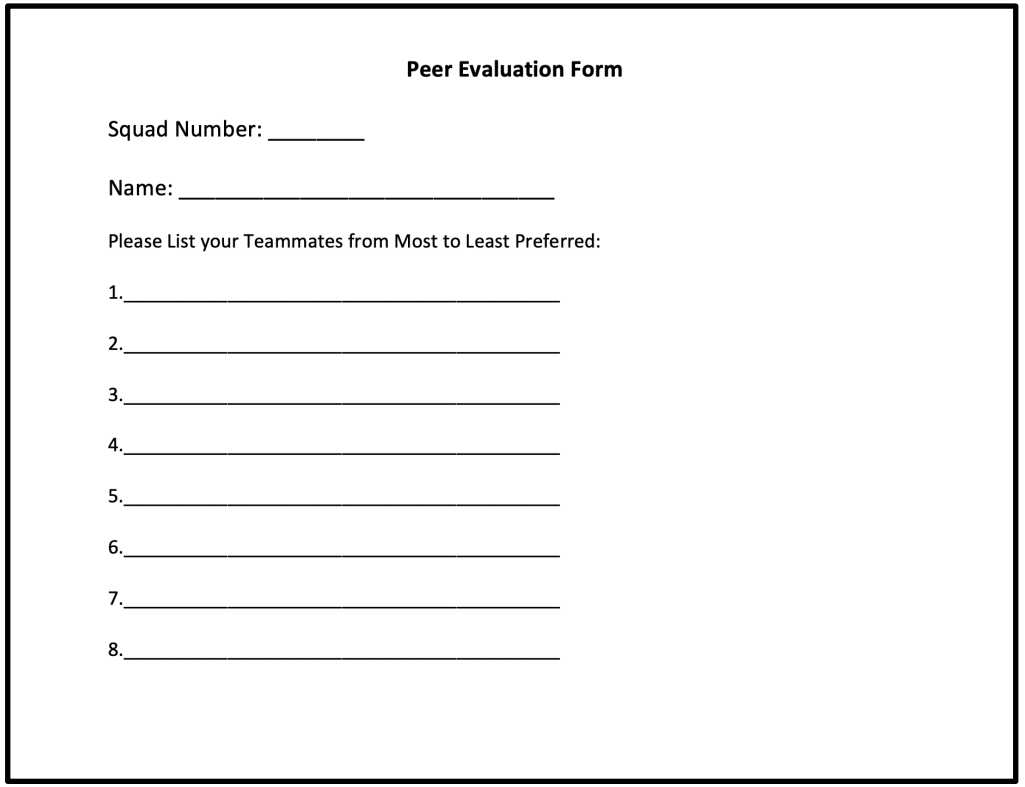
Summary

- Peer ratings are a critical component of selection and assessment
- Peer ratings are often performed using a “rank order list” method
  - Collecting ranked lists is known to be psychologically unreliable
  - Mathematical methods to combine ranked lists are flawed
- Comparing pairs of teammates solves the data collection problems
- Processing pairwise comparisons with Elo scoring solves the mathematical problems
- The EloRater app provides an easy interface to conduct peer evaluations

Ranked Lists: Psychologically Flawed

- *The Magic Number 7, Plus or Minus Two* (1956) established that human beings cannot effectively compare more than 7 elements at one time
- This limitation is well-accepted and has driven the course of Decision Sciences research since the 1950's
- Pairwise Comparison is an often-used substitute, as seen in the popular Analytic Hierarchy Process (AHP)

USMA's (old) Summer Training Peer Evaluation



Peer Evaluation Form

Squad Number: _____

Name: _____

Please List your Teammates from Most to Least Preferred:

1._____

2._____

3._____

4._____

5._____

6._____

7._____

8._____

Elo Scoring

- Arpad Elo came up with the method in the 1950's to rank chess players
- All players start with the same number of points, each match is a zero-sum exchange of points from the loser to the winner
- Expected result of a match is based on the current score of each player

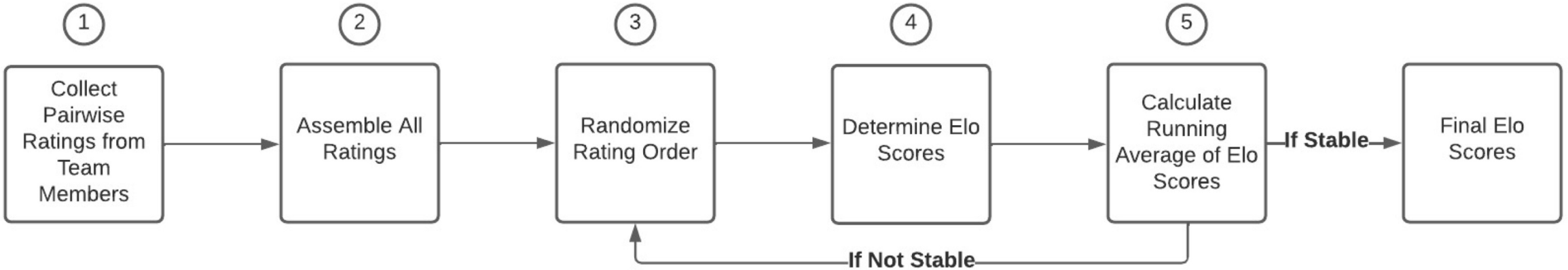
$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}$$

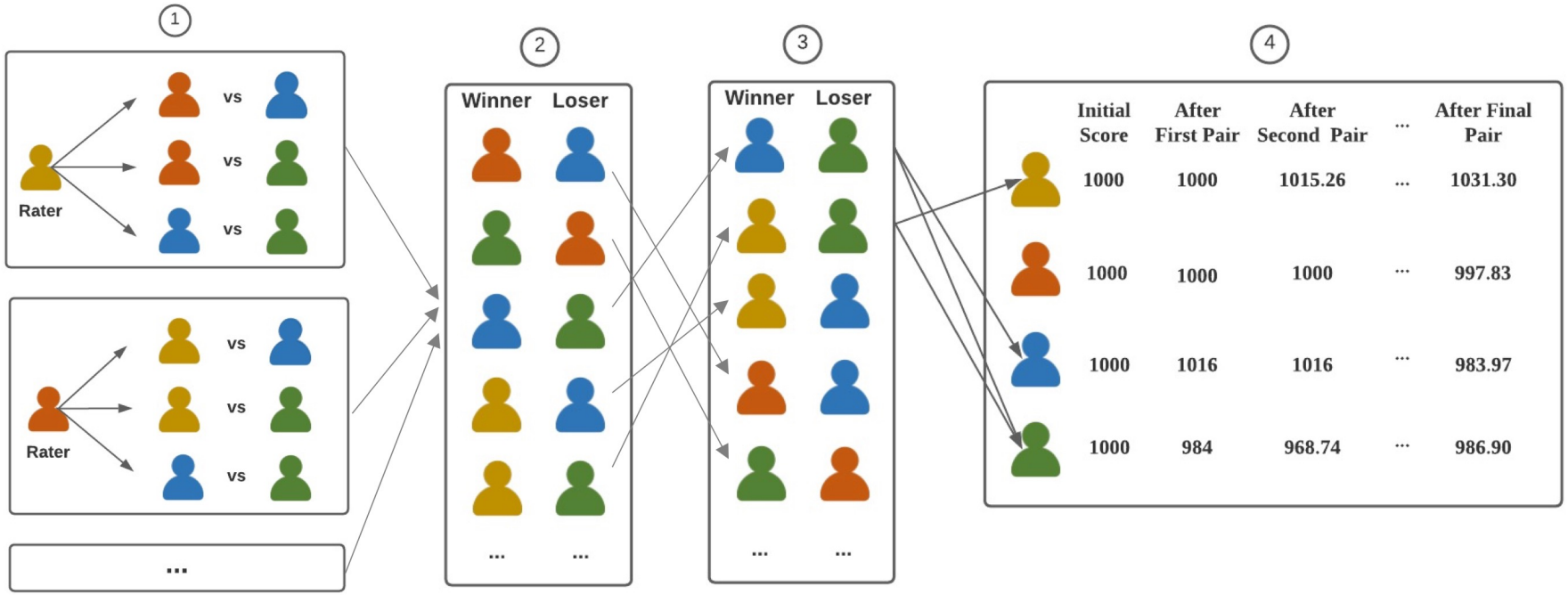
- Final scores are calculated with:

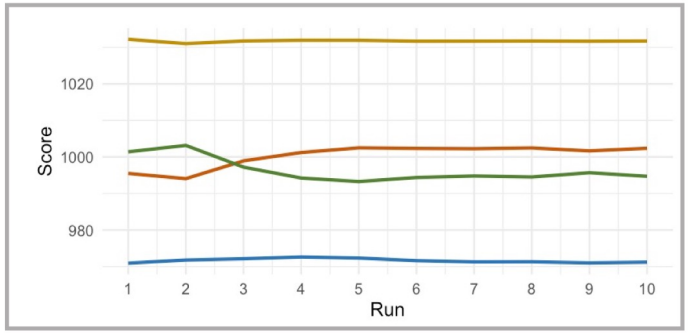
$$R'_A = R_A + K * (S_A - E_A)$$

where K is a constant and S is 1 if player A won, 0 if they lost

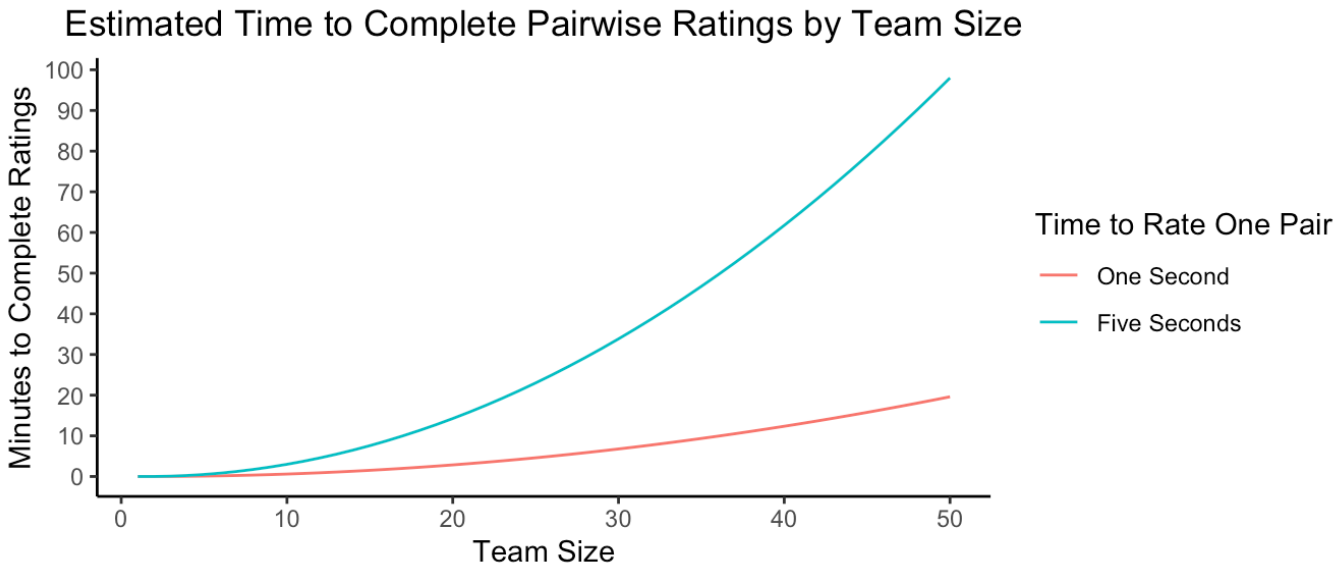
An “Elo Scoring” Method for Combining Comparisons into Consensus Scores







Pairwise/Elo Method: Practicality and Limitations



- The number of pairwise ratings we need to collect depends on the number of people on the team
- Teams with <30 members create reasonable completion times
- With larger teams, it is possible to collect incomplete data (i.e., users only rate a subset of comparisons) while still arriving at a reliable team consensus

EloRater: An App for Collecting and Processing Peer Evaluations

- All software is open-source
  - Python/Django
  - Anyone can easily add to the project
  - Full documentation on Github page

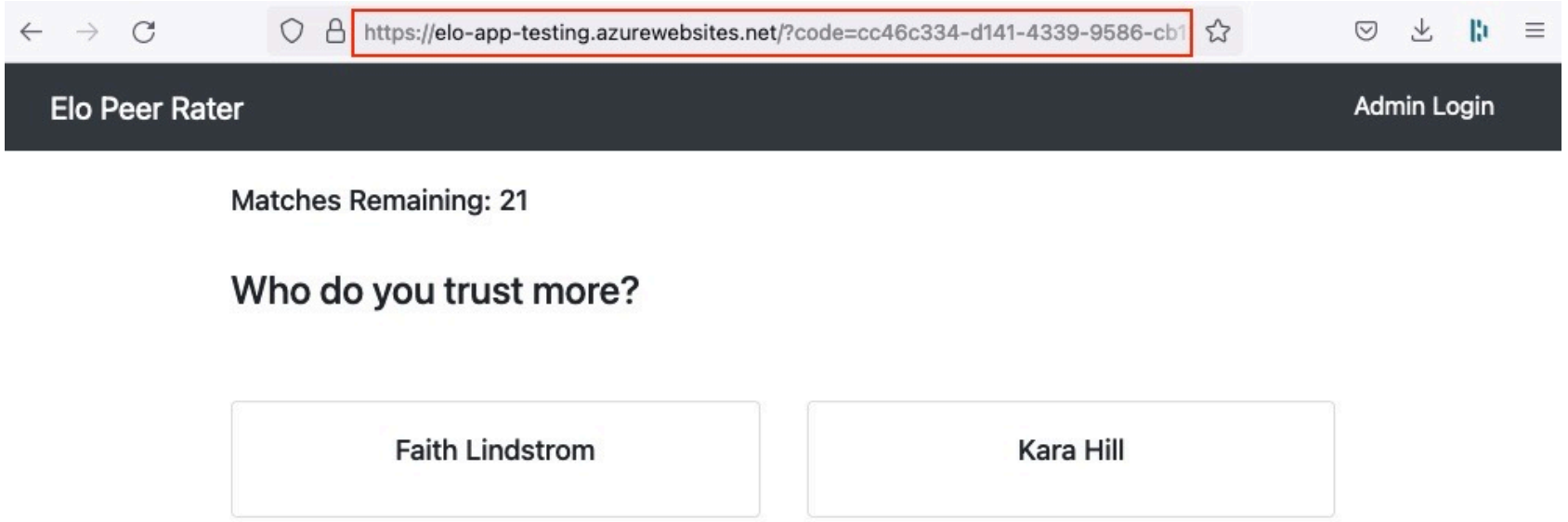
- Adaptable and Flexible
  - Current version has basic functionality
  - Easy to extend for specific use-cases (e.g., want to require free text comments)

Next Steps

- Continue App development to support specific use cases
- Share what works between active users/organizations
- Continue research/testing to ensure we are using the most defensible methodology that provides the most useful results
- Establish a hosted version on user-specific networks

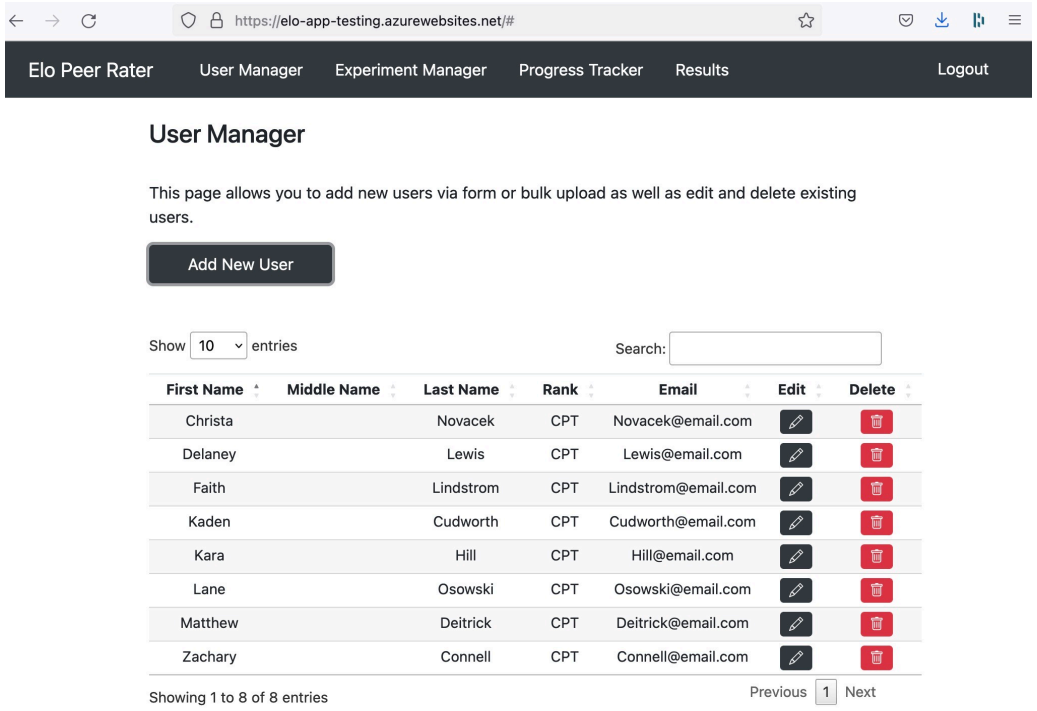
Questions?

BACKUP - EloRater: An App for Collecting and Processing Peer Evaluations

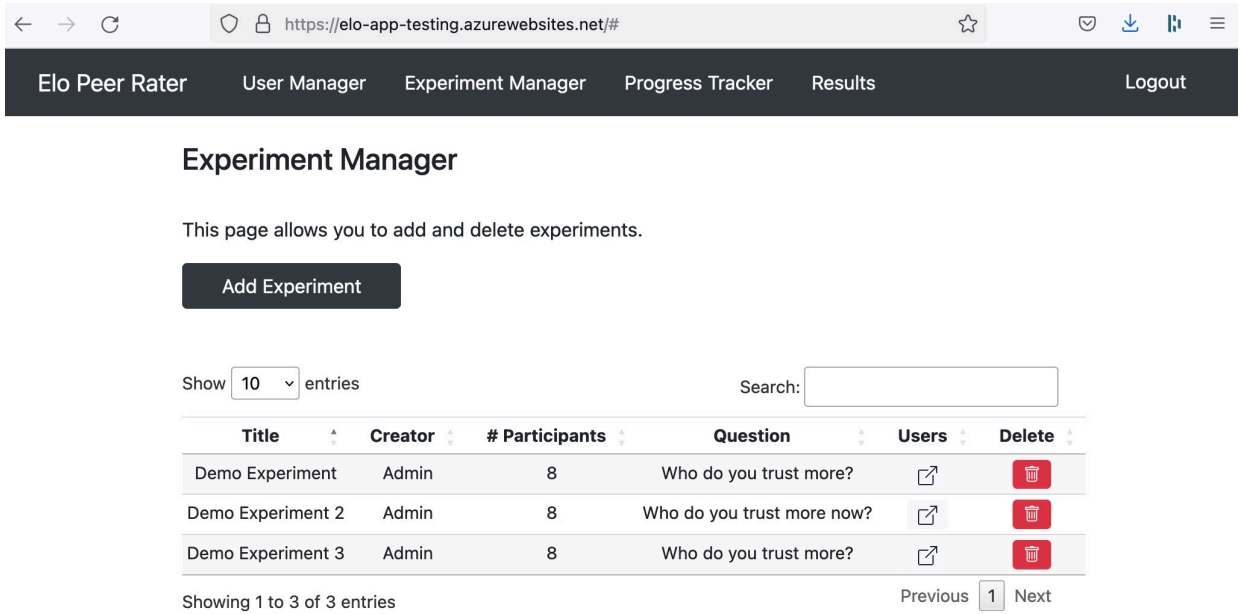


User View

BACKUP - EloRater: An App for Collecting and Processing Peer Evaluations

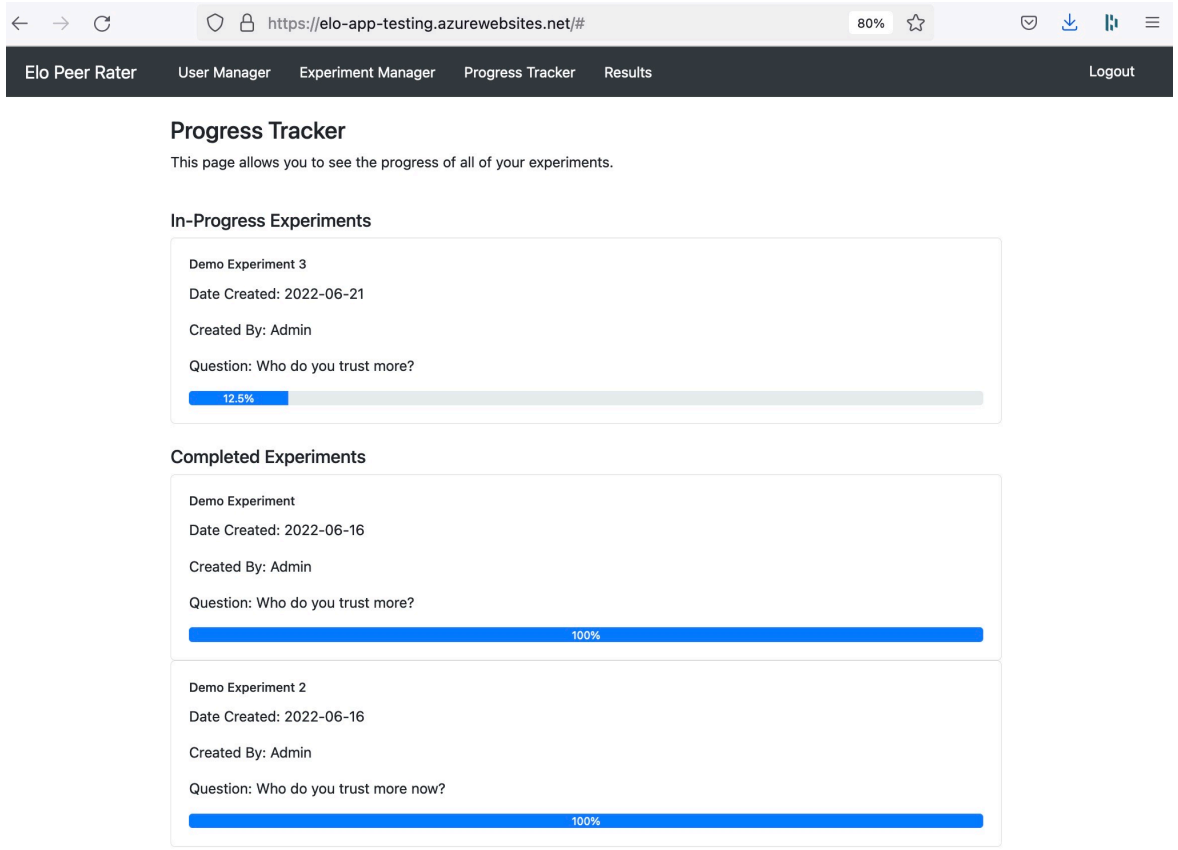


Adding People

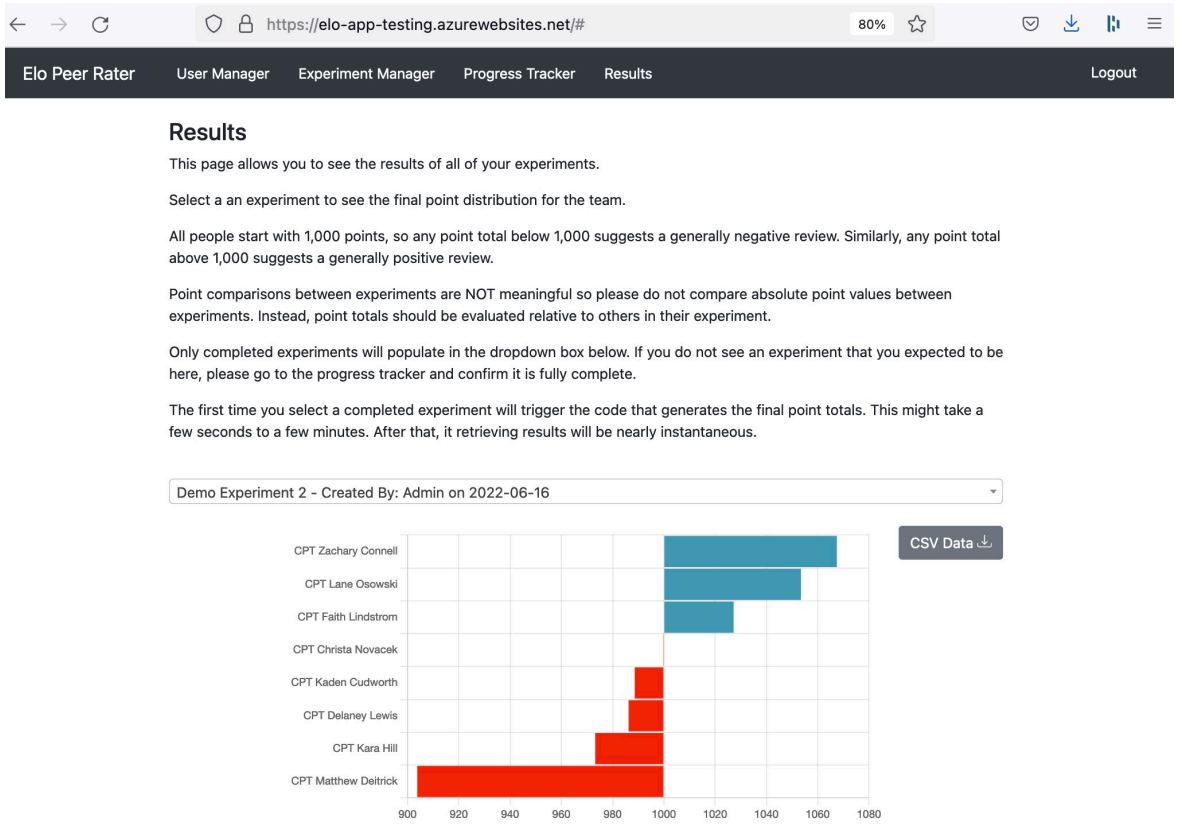


Adding Experiments

BACKUP - EloRater: An App for Collecting and Processing Peer Evaluations

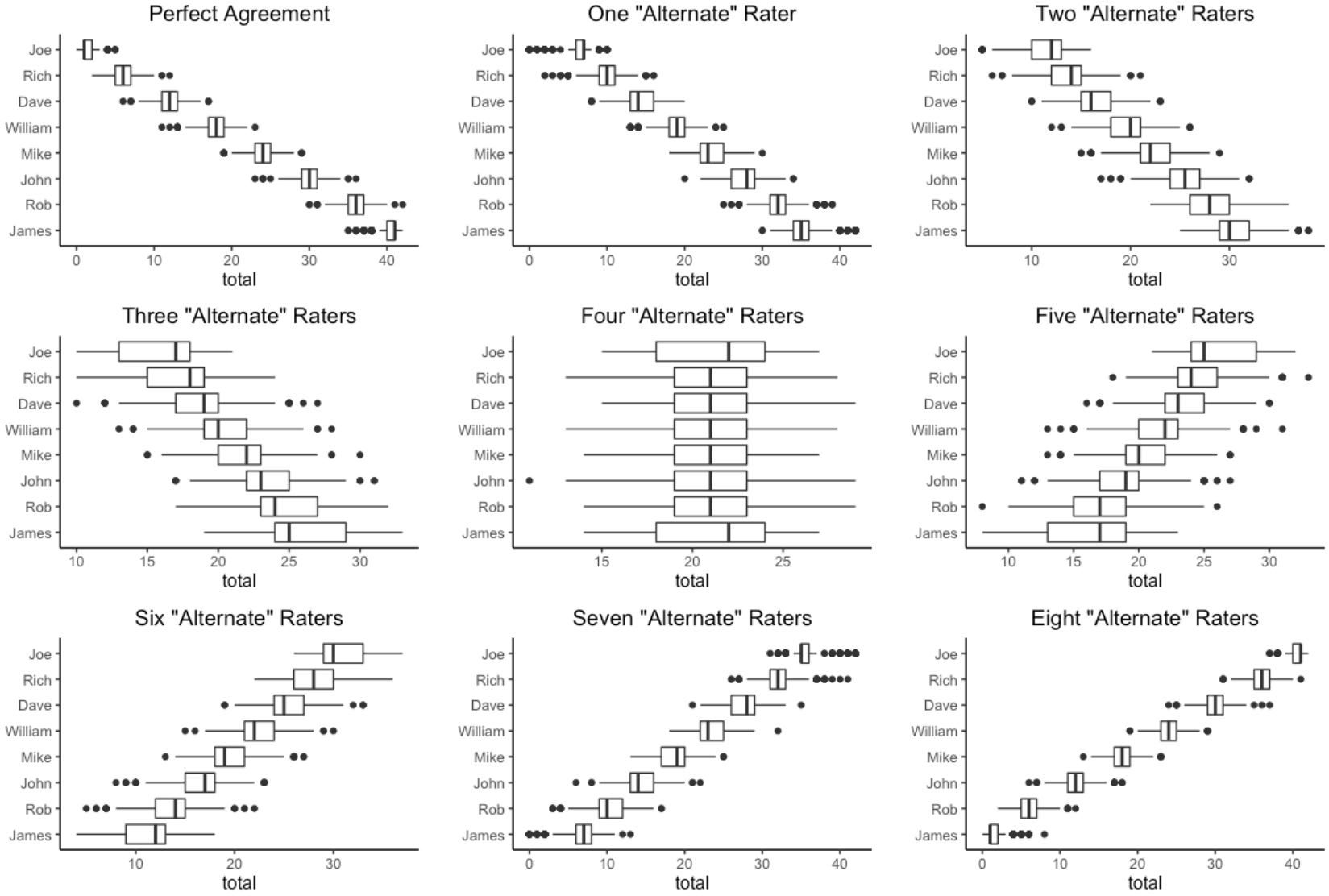


Tracking Progress



Viewing Results

BACKUP - Ranked Lists: Dlw?



BACKUP – Pairwise/Elo Method