

# 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

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

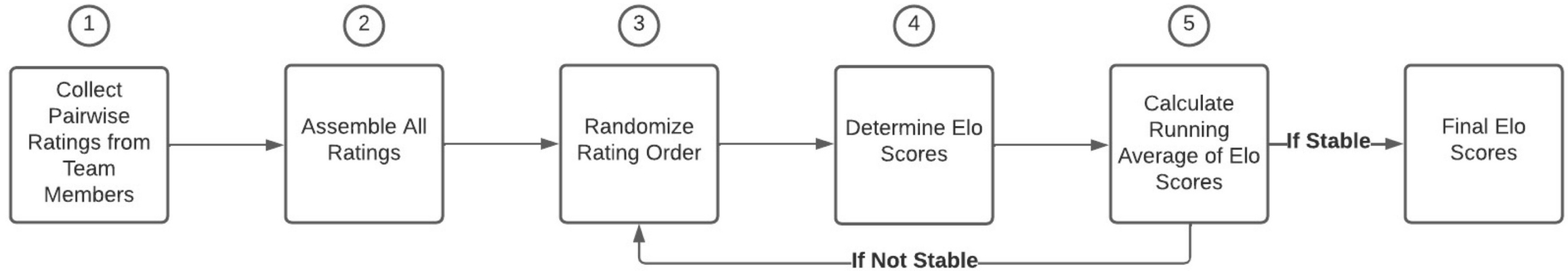
- Final scores are calculated with:

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

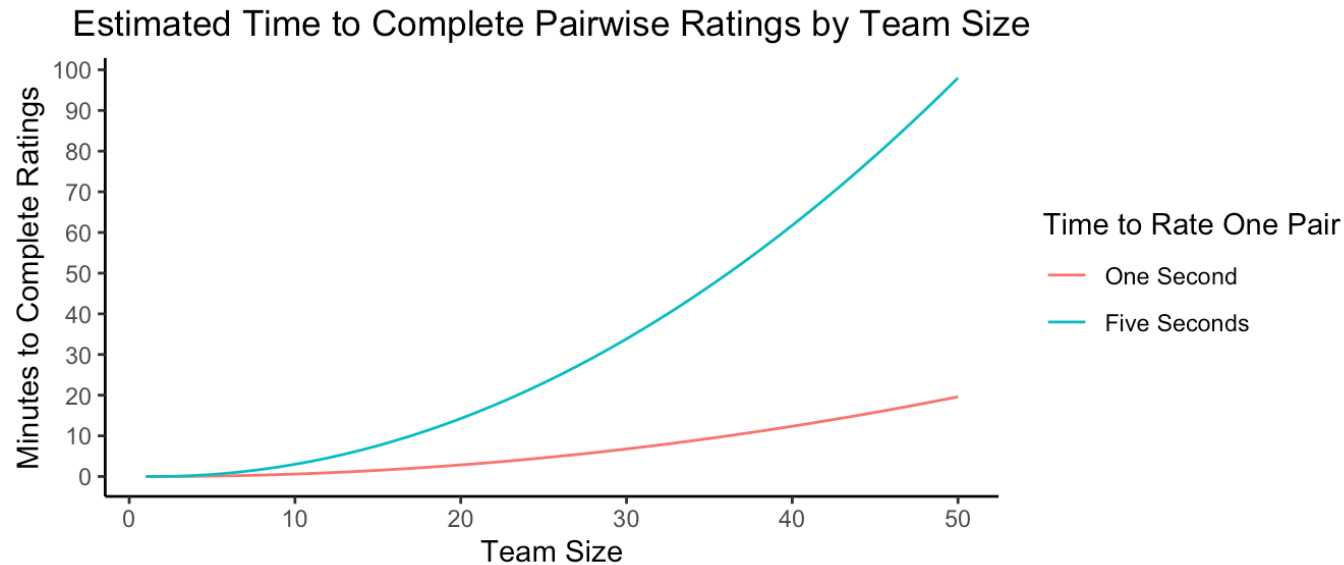
where K is a constant and S is 1 if player A won, 0 if they lost



# 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

---

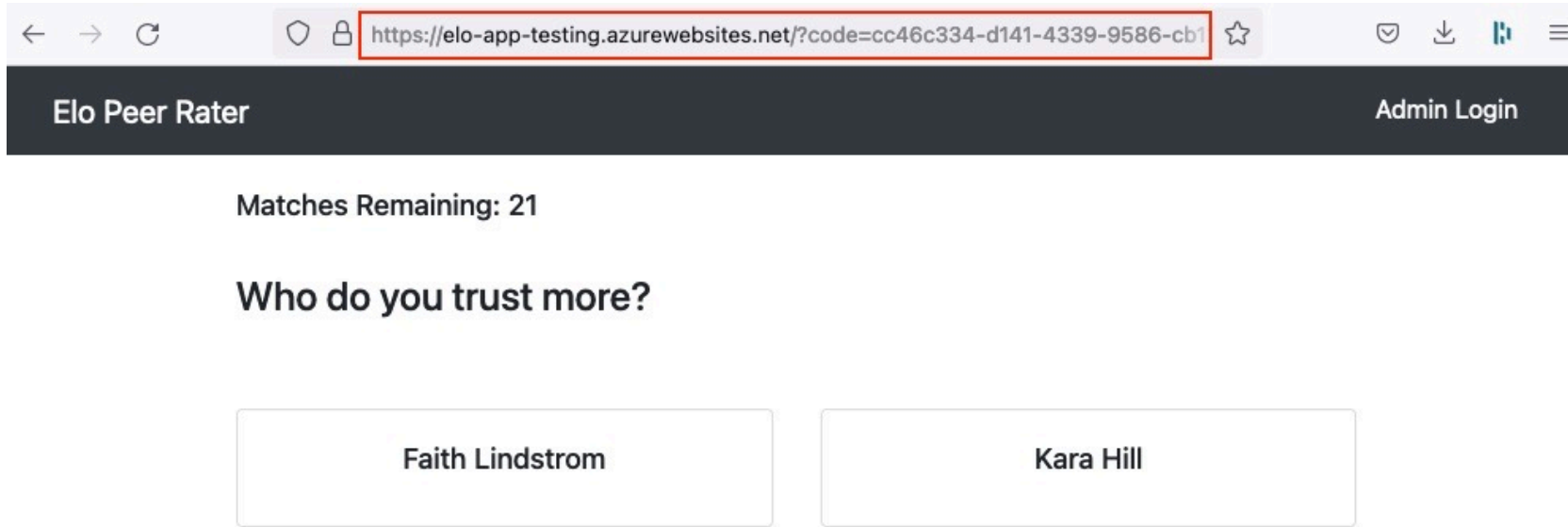
- Continue App development to support specific use cases
- Share what works between active users/organizations
- Continue research/testing to ensure we are using the most defensible methodology that provides the most useful results
- Establish a hosted version on user-specific networks



# Questions?

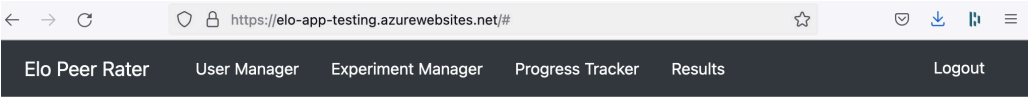
---

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



**User View**

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



## User Manager

This page allows you to add new users via form or bulk upload as well as edit and delete existing users.

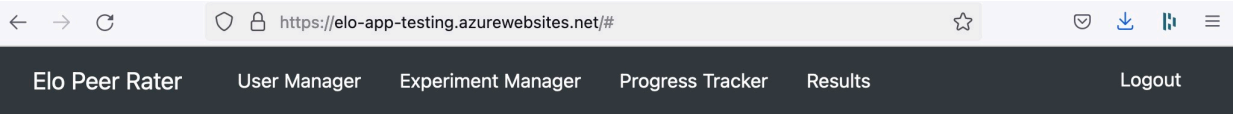
Add New User

Show 10 entries      Search:

First Name	Middle Name	Last Name	Rank	Email	Edit	Delete
Christa		Novacek	CPT	Novacek@email.com		
Delaney		Lewis	CPT	Lewis@email.com		
Faith		Lindstrom	CPT	Lindstrom@email.com		
Kaden		Cudworth	CPT	Cudworth@email.com		
Kara		Hill	CPT	Hill@email.com		
Lane		Osowski	CPT	Osowski@email.com		
Matthew		Deitrick	CPT	Deitrick@email.com		
Zachary		Connell	CPT	Connell@email.com		

Showing 1 to 8 of 8 entries      Previous 1 Next

## Adding People



## Experiment Manager

This page allows you to add and delete experiments.

Add Experiment

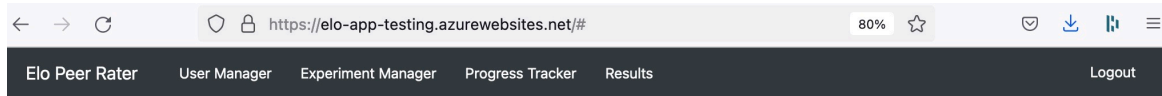
Show 10 entries      Search:

Title	Creator	# Participants	Question	Users	Delete
Demo Experiment	Admin	8	Who do you trust more?		
Demo Experiment 2	Admin	8	Who do you trust more now?		
Demo Experiment 3	Admin	8	Who do you trust more?		

Showing 1 to 3 of 3 entries      Previous 1 Next

## Adding Experiments

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



## Progress Tracker

This page allows you to see the progress of all of your experiments.

### In-Progress Experiments

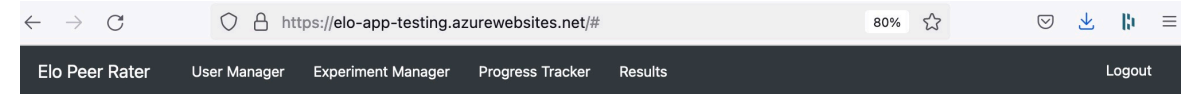
Demo Experiment 3  
Date Created: 2022-06-21  
Created By: Admin  
Question: Who do you trust more?  
12.5%

### Completed Experiments

Demo Experiment  
Date Created: 2022-06-16  
Created By: Admin  
Question: Who do you trust more?  
100%

Demo Experiment 2  
Date Created: 2022-06-16  
Created By: Admin  
Question: Who do you trust more now?  
100%

## Tracking Progress



## Results

This page allows you to see the results of all of your experiments.

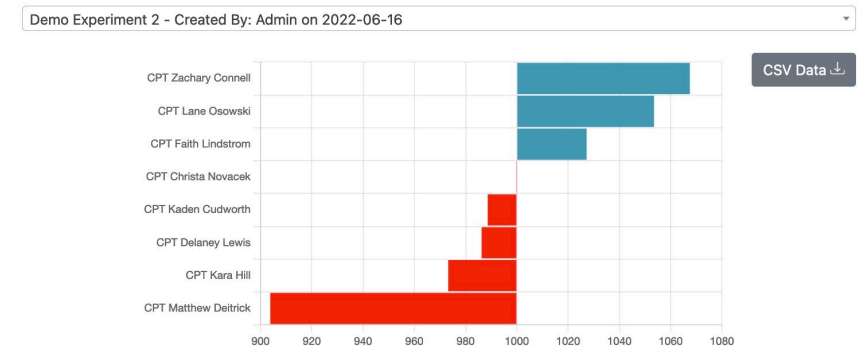
Select an experiment to see the final point distribution for the team.

All people start with 1,000 points, so any point total below 1,000 suggests a generally negative review. Similarly, any point total above 1,000 suggests a generally positive review.

Point comparisons between experiments are NOT meaningful so please do not compare absolute point values between experiments. Instead, point totals should be evaluated relative to others in their experiment.

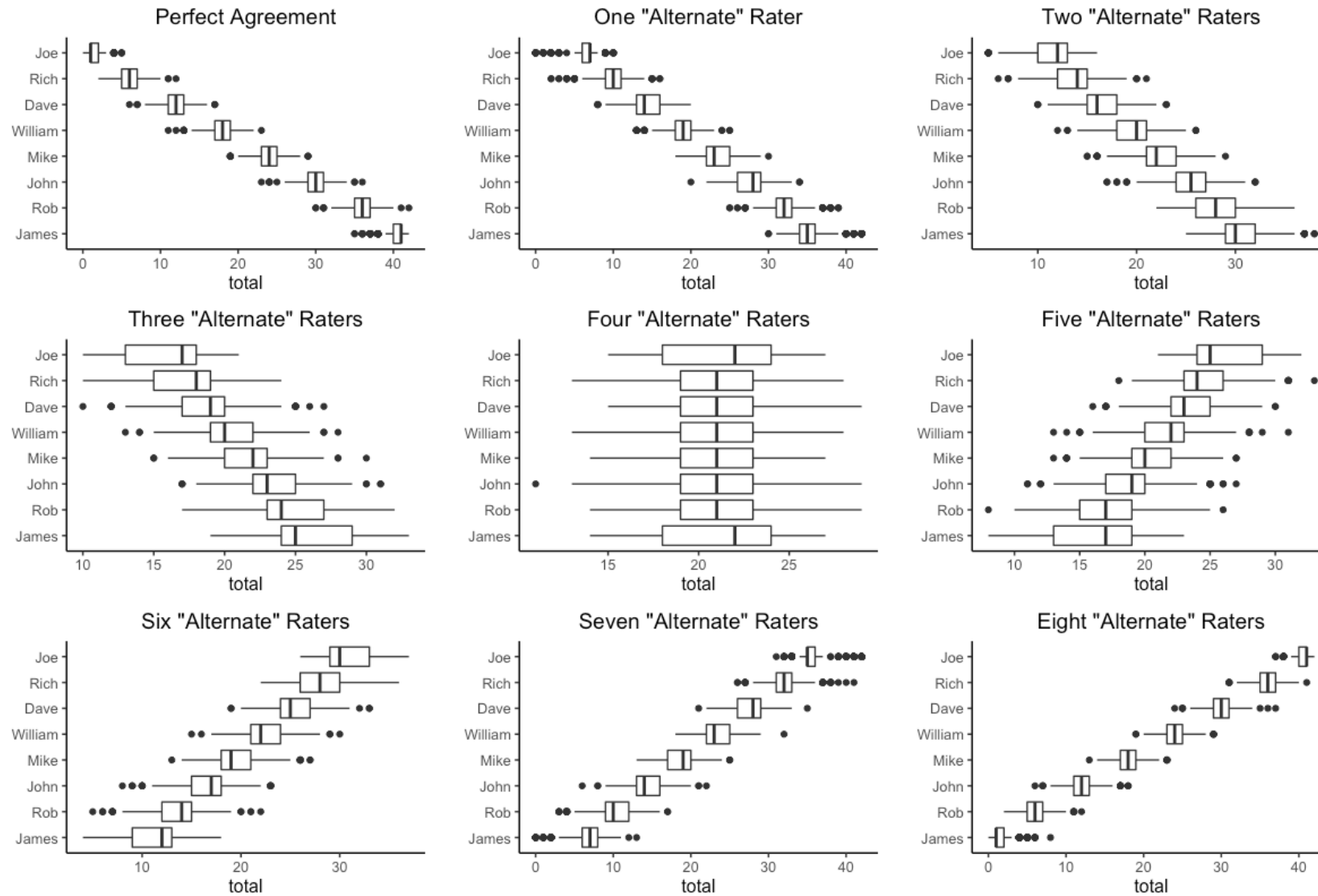
Only completed experiments will populate in the dropdown box below. If you do not see an experiment that you expected to be here, please go to the progress tracker and confirm it is fully complete.

The first time you select a completed experiment will trigger the code that generates the final point totals. This might take a few seconds to a few minutes. After that, it retrieving results will be nearly instantaneous.



## Viewing Results

# BACKUP - Ranked Lists: DIw?



# BACKUP – Pairwise/Elo Method

