Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 28, 2022
A Regional Conference of the Society for Industrial and Systems Engineering

Analysis of Peer Ranking Methods for Military Personnel Assessment

Marshall Pratt, Coert Pease, Zachary Glenn, Sijun Hwang, and Steven Lopez

Department of Systems Engineering
United States Military Academy
West Point, NY 10996

Corresponding author's Email: Coert.Pease@westpoint.edu

Author Note: Our team would like to thank Mr. Ian Kloo for his guidance and support throughout this research. We would also like to thank MAJ John Case for challenging our group to dive deeper into our understanding of peer rankings.

Abstract: Peer ranking is a commonly used method to gather information about relationships within a team. This information is often used in academic settings to assign grades, in corporate settings to build strong work units, and in personnel selection to determine the best person for a role. This paper will focus on the last use case, and more specifically how peer ranking can be used as a part of military personnel evaluation and selection processes. Although regularly used in this domain, there is little existing research regarding the suitability of different peer ranking methods for personnel selection. This paper provides an analysis of two popular peer ranking methods that are currently used in military selection: rank order listing and pairwise comparison. Using a simulation-based methodology, we demonstrate that both methods produce reliable results under varying conditions; however, we recommend using pairwise comparison due to well-known cognitive limitations with rank order listing.

Keywords: Peer Ranking, Peer Rating, Pairwise Comparison, Rank Order, Elo, Borda, Personnel Assessment

1. Introduction

Selection processes in the military use peer assessments to determine group dynamics and personal characteristics that are not easily seen from an observational position of authority. This process asks participants to, "judge the extent to which each of their fellow group members has exhibited specified traits, behaviors, or achievements," (Kane & Lawler, 1978). These peer assessments have been used in studies focused on leadership performance (Allen et. al, 2014), success in training (Zazanis et. al, 2001), and combat effectiveness (Williams and Leavitt, 1947). Through discussions with military practitioners (who asked not to be cited in this work), we found two commonly used peer assessment methods: rank order listing (where participants made lists of their teammates in order of preference) and pairwise comparison (where participants compared each of their teammates). Common reasons for using one method over the other were personal preference and institutional habits. While peer ranking is a well-researched topic in decision sciences, we found little existing work addressing peer evaluation in military selection, and we found no previous work investigating the validity of the pairwise method. This paper will first explore the existing research on peer assessment and the ways rank order listing and pairwise comparison are used in other domains. We will then describe a simulation-based methodology to evaluate the validity of both methods using known team preferences. Finally, we will recommend which method is most useful for military personnel selection and outline useful directions for future work in the problem space.

2. Background and Literature Reviews

Before describing our own experiment, this paper will explore existing research on peer assessment, as well as uses of rank order listing and pairwise comparisons in other areas of decision sciences. We narrow our focus of rank order listing to applications that employ Borda counting which is a popular method for combining rank order lists into a single, unified score. Similarly, we focus our study of pairwise comparisons on Elo scoring which serves the same function for pairwise comparison data. Our team selected these ranking methods due to their applicability in a personnel ranking situation.

2.1 Ranking Order Listing

Borda counting is a popular method for aggregating individual ranked lists into a unified team ranking. Named for Jean-Charles De Borda who developed the method in post French-Revolution Paris in 1781, Borda counting is designed to aggregate votes in a multi-decision maker context (Costa, 2017). The goal of Borda counting is to formulate a combined-global ranking from multiple candidate rankings with the least damage to consistency (Saari, 1985). The Borda counting process starts with each decisionmaker producing a list ranking a set of candidates in order of preference. These rankings are then converted into scores by taking the number of team members (n) and subtracting the rank. For example, the most preferred candidate receives a score of n-1, the second most preferred a score of n-2, and so on. Each candidate's scores are summed and the candidate with the most points is deemed the winner (Felsenthal, 1996).

While the Borda is simple and reliable, De Borda himself notes that his method was not impervious to the ill-intentioned voter, claiming "[his] scheme is only intended for honest men," (Felsenthal, 1996). For example, if a voter (or decision maker) knew their top candidate would be in close competition with another, they may be inclined to place that candidate far lower than their true preferences so that their top candidate received the most points and their candidate's competition received none. This poses obvious issues when looking to create an accurate global ranking from each voter's ballot and introduces a fundamental contradiction to any method looking to combine preferences of multiple decision makers into one (Arrow, 1950). Certainly, any peer evaluation system is vulnerable to insincere raters, but Borda counting makes this gamesmanship particularly easy to accomplish. Even with well-intentioned raters, however, Borda is still susceptible to a major problem that stems from the fact that the input to the system is rank ordered lists.

2.1.1 The Problem with Rank Order Listing

The process of formulating a rank ordered list of preferences is not a simple task for human cognition. Thomas L. Saaty highlighted this limitation in his study, *Why the Magic Number Seven Plus or Minus Two*, where he determined that seven preferences (plus or minus two) were too many for humans to accurately and reliably rank (Saaty, 2003). He found that after roughly seven elements were compared, humans became desensitized to the differences between them and therefore could not make accurate judgments (Saaty, 2003). This evidence suggests that, though it is often used for peer assessments, rank order listing is not advised with teams with more than four members.

2.2 Pairwise Comparison

While there are many different methodologies for evaluating pairwise rankings, we found that the Elo rating system is particularly well suited for use in peer assessments. Developed by Arpad Elo, "The Elo-rating calculation procedure is based on the assumption that the chance of A winning from B is a function of the difference in current ratings of the two contestants," (Albers and Vries, 2000). For example, a player with a higher Elo score is *expected* to defeat a player with a lower score. If the higher ranked player wins, fewer points are exchanged between the players. If, however, the lower ranked player wins, a larger number of points are exchanged. Equation 1 is used to determine the expected score of a game between Players A and B. The variables R_A and R_B are the ratings of each player respectively. E is a continuous number running between 0 and 1. Player A is expected to win overall if E is 1, tied if it is $\frac{1}{2}$, and lose if it is 0 (Glickman and Jones, n.d.).

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

Equation 2 is used to determine each player's updated Elo score after the match is played. R_{post} is a player's updated rating, while r_{pre} is their rating before the match. K is a constant factor that is pre-determined, defining the magnitude of how much a player's rating can change in one match. This constant k is dependent on the number of matches played, which is a larger value for newer players and a smaller value for older players (Albers and Vries, 2000). This allows for newer players to quickly approach an appropriate rank that reflects one's skill level, while older players have a slower repositioning once their Elo score reflects their skill level (Albers and Vries, 2000). $S - S_{exp}$ is the difference between the player's score, and the player's expected score, which is "E" from Figure 1 (Glickman and Jones, n.d.).

$$r_{post} = r_{pre} + K(S - S_{exp}) \tag{2}$$

Although Elo scoring is not widely used in peer assessment, it is a staple of soccer analytics (Wolf et al., 2020) and E-Sports (Ebtekar, Aram., Liu, Paul., (2021). To adapt the Elo methodology for peer assessment, each member of a team is asked to rate all possible pairwise comparisons of their teammates.

3. Methodology

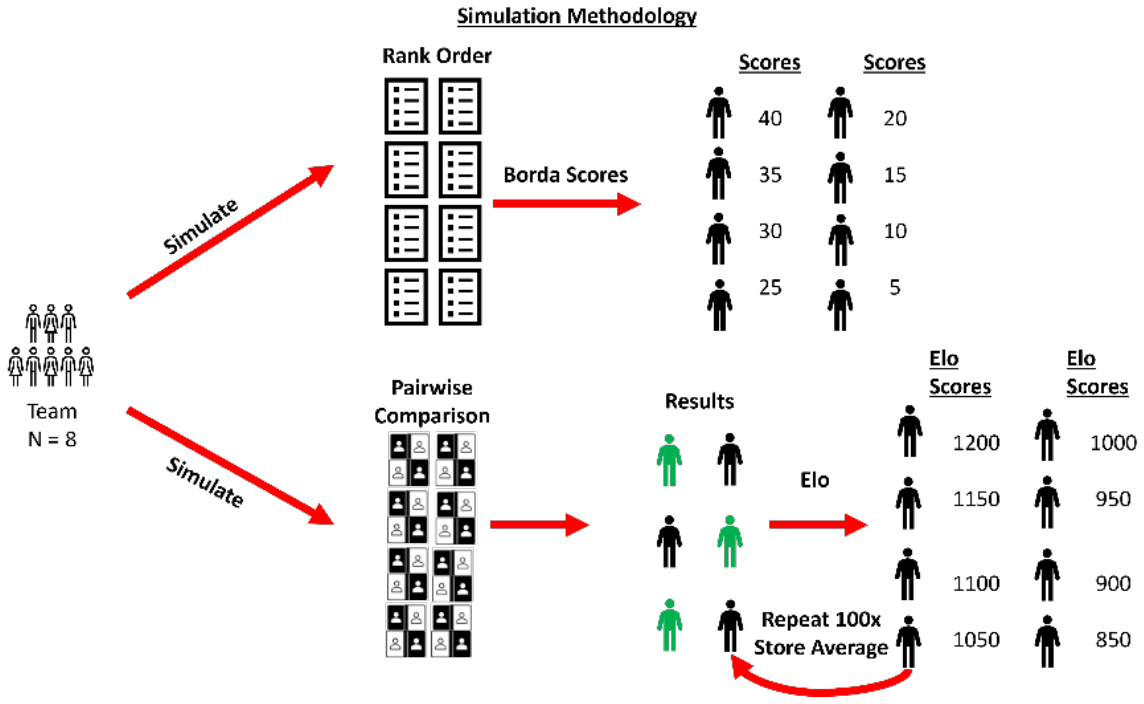

Figure 1. Graphic of Borda and Elo Ranking Methods

While there appears to be support for using both pairwise comparison and rank order listing in existing literature, we did not find any studies evaluating their effectiveness in realistic team settings. Although we had access to real world peer assessment data, we decided to simulate data to have a controlled list. We did this because it is impossible to know the true rank order of the real-world data, therefore it would not be a helpful metric to compare Elo or Borda rankings. Our simulation allowed us to test if both the Borda and Elo methods provide reliable results when we knew a team's exact preferences. Our methodology involved generating rank order and pairwise data under various levels of team consensus, processing that data using Borda and Elo methods (respectively), and finally evaluating the consistency of the output compared to the known team preferences.

3.1 Simulation

We simulated our data using teams of size 8 – roughly a squad sized element – with varying levels of team consensus (i.e., the number of team members with identical preferences). When creating the rank order data, each team member generated a preference list starting from the top of their preferences and moving to the bottom. At every rank position, there is an 80% chance the rater would select the highest ranked remaining person with a 20% chance of selecting the next highest ranked alternative. The percentages assigned for the positions are a rough attempt at replicating reality. This synthetic error is meant to make the simulation closer to reality since individuals often have unclear preferences for similarly ranked choices. The pairwise data was also generated such that each rater selected the top-ranked alternative 80% of the time.

3.2 Scoring

After generating the rank order data, creating the Borda scores was done so according to the process described in Section 2.1 of this paper. The Elo rankings required a more nuanced approach considering the process' iterative nature. Because the participants' Elo rankings entering a "match" determine the number of points exchanged (see section 2.2), the order that the pairs are adjudicated when creating the Elo scores is important. To remove this time effect, we reordered the pairwise results and calculated Elo scores 100 times for each simulation run, storing the running average of the scores. After this process, the output can be evaluated like the Borda scores where more-preferred preferences get higher scores.

3.3 Iteration

We iterated the simulation process 100 times under four different conditions. First, we allowed all team members to have identical preferences. Then, we introduced team members with the opposite preference one at a time until three of the team members disagreed with the original five. We call these raters with opposite preferences "bad raters" to note that their ratings are the opposite of percentages assigned for "normal raters." In practice, a good peer assessment metric should reflect a declining consensus among the group with the increased number of "bad raters."

4. Results and Discussion

Figure 2 shows the distributions of Borda scores that each team member received over 100 simulation runs. The plot labeled "No 'Bad' Raters" was generated using complete team consensus and it shows clear separation between the scores between the team members, suggesting the method can reliably show the team's true preferences. The variation in Borda scores in this plot is the result of the imperfection built into the rank order simulation. Even with poor rating capability (20% chance of picking the wrong person), the results are stable. As we introduce "Bad Raters" (i.e., raters with differing preferences), the simulation output accurately shows the degradation in the team's consensus, as evidenced by the increasingly overlapping score distributions. This suggests that the Borda method can combine individual rank order lists into a unified team ranking in a way that accurately reflects the teams' true preferences.

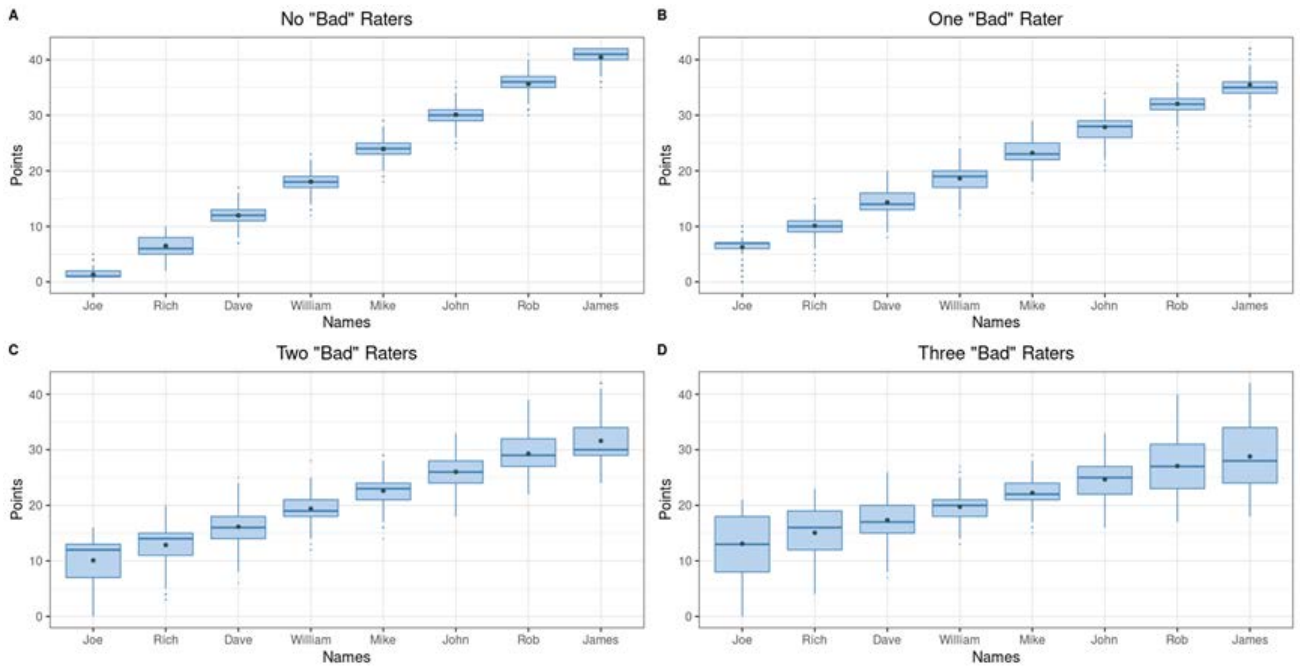

Figure 2. Plots indicating the distribution of Borda scores with varying levels of team consensus

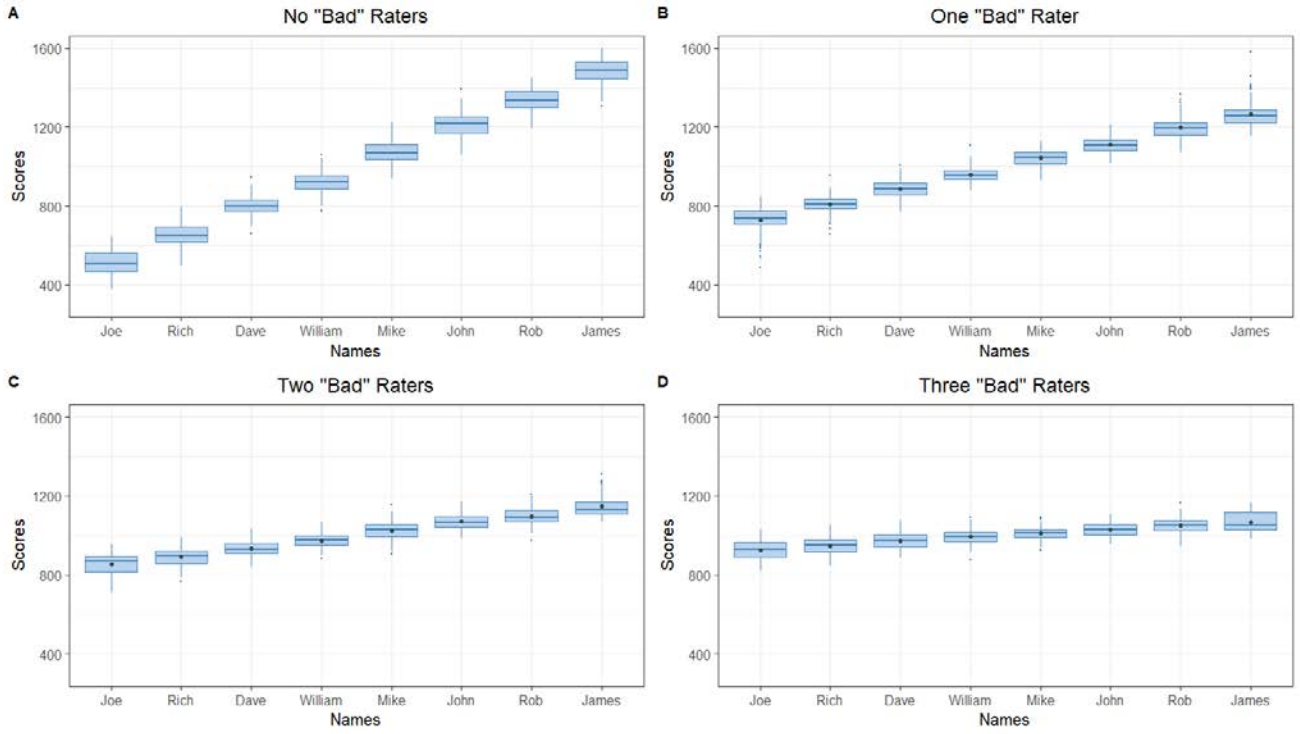

Figure 3. Plots indicating the average distribution of ranks for Elo with increasing number of incorrect raters

Similarly, Figure 3 shows the resulting Elo scores from 100 simulation runs for each team member. As with the Borda scores, there is clear distinction between the team members when the team's preferences are aligned. As the team's preferences are made murkier (by introducing contradictory raters), the output reflects this change with increasingly overlapping distributions. Interestingly, unlike with the Borda method, the Elo distributions do not appear to widen as the team consensus degrades. This is potentially an artifact of the way the rank order lists were generated in the simulation and warrants further examination (more discussion on this below in Section 5). Overall, as with Borda counting, Elo scoring appears to be a valid method for conducting peer assessments under the conditions of our simulation.

5. Limitations and Future Work

While our study demonstrated that both Borda and Elo methods can be used to produce reliable peer assessments under known conditions, there is an urgent need for more research in this space. A logical first step would be to introduce more realistic parameters in the simulation. For example, it would be useful to study how people fill out rank order lists (e.g., top-down, bottom-up, or something else entirely) to create a more accurate simulation. We advised against extrapolating too many conclusions from the widening distributions in the Borda distributions in Section 5 because we are unsure if the top-down method for generating rank orders is actually how people complete these lists. Further study into this generating process will enable more robust analysis of these results. Similarly, it is possible that there are intricacies in the ways that people approach pairwise comparisons that we did not consider.

Beyond the potential unintended problems with the nature in which people generate rank order lists, we identified that the Borda method is vulnerable to gamification that has been cautioned against since the method was first devised. We suspect that pairwise comparisons are unlikely to be gamed effectively since it would require participants to memorize an unreasonable amount of information. These human factors were not considered by this study but are important to identify potential weaknesses in both methods. This study could also be extended to demonstrate the effect of team size and mixing. In particular, the Elo method relies on pairwise comparisons which would grow unwieldy with large group sizes. It is also possible, however, the cognitive difficulties we identified with rank order listing could be exaggerated as team sizes grow. Some of these problems

could be solved by gathering incomplete information (i.e., only asking raters to rate a subset of their peers), but this would require careful study into how much information is required to give reliable results.

6. Conclusions

After a thorough literature and simulation study evaluating rank order listing and pairwise methods for peer assessment, we strongly recommend the pairwise method (using Elo scoring) over the rank order method for military assessment. While our simulations demonstrated that both the Borda (rank order) and Elo (pairwise) methodologies provided reliable rankings given known team preferences, our research into human cognition suggests that we cannot expect raters to provide consistent data with teams larger than four people. Many important research questions remain to determine the best way to execute pairwise comparison in a team setting, but it appears that the Elo method we presented here works well under reasonable conditions. There are many future lines of research that could strengthen this study's conclusions. The authors of this paper hope the simulation methodology presented in this paper serves as a point of departure for this important future work.

7. References

Albers, Paul C.H., and Vries, Han de. (2000). Elo-Rating as a Tool in the Sequential Estimation of Dominance Strengths. *Animal Behavior*, 2001. Accessed 21 March 2022. https://idealibrary.com

Allen, M. T., Bynum, B. H., Oliver, J. T., Russell, T. L., Young, M. C., & Babin, N. E. (2014). Predicting leadership performance and potential in the U.S. Army Officer Candidate School (OCS). *Military Psychology*, 26(4), 310–326. https://doi.org/10.1037/mil0000056

Arrow, Kenneth. J. (1950, August). A Difficulty in the Concept of Social Welfare. *The University of Chicago Press - Journal of Political Economy*. Vol. 58, No. 4, pp. 328-346. https://www.jstor.org/stable/pdf/1828886.pdf

Boyd, Andrew. (2003) Arrow's Paradox. *The Engines of Our Ingenuity. Houston, Texas: The University of Houston – Oxford University Press*. https://www.uh.edu/engines/epi2427.htm

Costa, Helder. G.(2017). *AHP-De Borda:* A Hybrid Multicriteria Ranking Method. *Brazilian Journal of Operation Production Management*, pp. 281-287. https://doi.org/10.14488/BJOPM.2017.v14.n3.a1

Ebtekar, Aram., and Liu, Paul., (2021). *Elo-MMR:* A Rating System for Massive Multiplayer Competitions. *International World Wide Web Conference Committee*. https://cs.stanford.edu/people/paulliu//files/www-2021-elor.pdf

Felsenthal, Dan. S. (1985). "Setting the Record Straight: A Note on Sophisticated Voting under Borda's Method". *JSTOR* https://www.jstor.org/stable/30024145

Glickman, Mark E., and Jones, Albyn C. (n.d.). *Rating the Chess Rating System*. Accessed December 10, 2021. http://www.glicko.net/research/chance.pdf

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological Bulletin*, 85(3), 555–586. https://doi.org/10.1037/0033-2909.85.3.555

Saari, Donald, G. (1985, January). The Optimal Ranking Method is the Borda Count. Northwestern University, *Kellog School of Management, Center for Mathematical Studies in Economics and Management Science*. Evanstan, IL.. https://www.econstor.e u/bitstream/10419/220997/1/cmsems-dp0638.pdf

Saaty, T.L and Ozdemir, M.S. (2003, July). Why the Magic Number Seven Plus or Minus Two. *Mathematical and Computer Modelling 38*. 233-244.

Williams, S. G., & Leavitt, H. J., (1947). Group opinion as a predictor of military leadership. *Journal of Consulting Psychology*, 283-291. https://doi.org/10.1037/h0056512

Wolf, Stephan., Schmitt, Maximilian., Schuller, Bjorn. (2020). A Football Player Rating System. *Journal of Sports Analytics 6*. https://content.iospress.com/download/journal-of-sports-analytics/jsa200411?id=journal-of-sports-analytics%2Fjsa200411

Zazanis, M. M., Zaccaro, S. J., & Kilcullen, R. N. (2001). Identifying motivation and interpersonal performance using peer evaluations. *Military Psychology*, 73. https://doi.org/10.1207/S15327876MP1302_01