Improved Peer Evaluations Using Elo Scoring and Sentiment Analysis

Nico Manzonelli, Jeff Mayolo, Samar Sikka and Joseph Zuccarelli

May 3, 2022

Peer Evaluations at West Point

Every summer cadets at the United States Military Academy, commonly known as West Point, conduct a series of military training exercises to develop their tactical expertise and leadership skills. Two core training exercises are Cadet Field Training (CFT)–completed prior to a cadet's 2^{nd} year at the Academy–and Cadet Leadership Development Training (CLDT)–completed prior to a cadet's 4^{th} year at the Academy. Throughout the duration of both these training exercises, cadets are placed in platoon and squad sized elements of 32 and 8 individuals, respectively. When summer training nears an end, each cadet is asked to evaluate their peers based on several factors such as tactical proficiency, team-working ability, and leadership potential.

Following the conclusion of the 2019 Cadet Summer Training (CST) cycle, a group of faculty researchers from the Department of Systems Engineering carried out a thorough analysis of the peer evaluation data collected over the summer. Under the existing peer evaluation system each cadet ranked their squad mates overall performance from 1 to $n-1$, n being the size of the squad. Note that cadets were asked not to rank themselves. With each ranking cadets were also required to provide a qualitative justification for their decision (see Appendix A for an example ranking card).

When evaluating the collection of disjointed rankings and qualitative assessments gathered at the conclusion of CST, academy instructor and data scientist Ian Kloo noticed a few issues presented by the data. First, Mr.Kloo observed that most group ratings were consistent in ranking the top and the bottom cadets. However, rankings for cadets in the middle of the squad were jumbled and seemed random. This made it difficult to infer the team-working abilities of the "middle of the pack" cadets. Furthermore, Mr. Kloo also found that the qualitative justifications provided by cadets alongside their rankings were not being used in an efficient manner. These justifications were only given any consideration when strong performers received low peer rankings or poor performers received high peer rankings. Given these findings, how could Mr. Kloo advise Academy leaders to adjust the peer evaluation system to provide improved ratings and make better use of qualitative assessments?

Why Peer Evaluations are Important

The US Army is an organization where team trust and cohesion are essential for success. One toxic team member can jeopardize the culture and performance of an entire unit. Therefore, the Army requires methods for evaluating an individual's ability to build relationships and work as part of a team. In order to do this, the Army relies heavily on peer evaluations. Although characteristics such as fitness level and tactical proficiency are important in terms of identifying quality soldiers, the Army seriously considers how each individual is ranked by their peers. For instance, soldiers in Ranger School, one of the Army's most rigorous training courses, can fail out of the course due to poor peer evaluations.

At West Point, the impact of peer evaluations is less amplified, yet still relevant. While bad peer evaluations are not grounds to remove a cadet from the Academy, it does impact cadets' grades during summer training. In addition to being graded on academic performance, cadets receive grades for their military performance. Military grades can severely impact a cadet's career trajectory, as they are factored into branch and unit selection prior to graduation. Beyond grades, peer evaluations impact the quality of leadership development. West Point's primary goal is to develop quality leaders for the US Army. Therefore, the qualitative and quantitative aspects of each cadet's performance are given careful consideration and reflection. Every cadet is assigned to a senior Army officer who mentors them throughout different stages of their cadet-career. After each summer training exercise, these officers conduct formal counseling sessions with their cadets in order to analyze and evaluate their overall performance, which includes peer evaluations.

Given the broad impact of peer evaluations on cadets, it is important to make them as accurate and fair as possible. Prior to 2019, the Army and West Point alike conducted peer evaluations using a traditional 1 to $n-1$ ranking approach where each member ranked the other members of the unit against each other and provided short qualitative assessments of other members in the form of "sustains" and "improves." This method of evaluation was useful in that it was very easy to implement; however, as discovered by Mr. Kloo, it failed to provide useful summary statistics and was prone to major bias. As an alternative, Mr. Kloo and his research team within the West Point Systems Engineering Department investigated implementing Elo scoring as a substitute for traditional rank-based peer evaluations.

Elo Scoring

Named after its creator, Arpad Elo–a Hungarian-American physics professor born in 1903, Elo scoring is a method for calculating the relative skill levels of players in zero-sum two-player games. Elo first developed this rating system in 1960 as an improved method for ranking chess players [1]. The United States Chess Foundation, which Elo participated in as a master-level chess player since its founding in 1939, served as the first organization to ever implement Elo scoring. Shortly after this, the scoring method gained widespread recognition as being fairer and more accurate than previous scoring systems, leading to its adoption by the World Chess Foundation in 1970 [2]. Since then, Elo scoring has been used in various other settings such as sports, video games, and dating apps [3].

Performance within an Elo scoring system is not measured in absolute terms. Rather, it is inferred from wins, losses, and draws against other players. Each individual's Elo score is dependant upon both the ratings of their opponents and the outcome of the match. After each game, the winning player gains points from the losing player. The amount of points involved in this transfer is determined by the difference in the two players' ratings. If the higher-rated player wins, a few points are gained from the lower-rated player. If the lower-rated player wins, a lot of points are gained from the higher-rated player. In the case of a draw, the lower-rated player gains a few points from the higher-rated player [4].

The mathematics underlying the scenarios presented above are as follows. The fundamental assumption behind Elo scoring is that each player possesses a current playing strength that is unknown, yet can be estimated by a rating. Consider two players, Player A and Player B. Let R_A represent the rating of Player A and let R_B represent the rating of Player B. The expected score of Player A, E_A, and the expected score of Player B, E_B, are calculated as follows:

$$\begin{aligned} E_A &= \frac{1}{1+10^{\frac{R_B-R_A}{400}}} \\ E_B &= \frac{1}{1+10^{\frac{R_A-R_B}{400}}}. \end{aligned}$$

Let S_A and S_B represent the points scored by Player A and Player B, respectively. Note that players score 1 point for winning a match, 0 points for losing a match, and 0.5 points for ending in a draw. Using the two equations defined above, the formulas for calculating Player A's updated rating, R'_A, and Player B's updated rating, R'_B, following the conclusion of a match are defined as follows:

$$\begin{aligned} R'_A &= R_A + K(S_A - E_A) \\ R'_B &= R_B + K(S_B - E_B). \end{aligned}$$

Note that the weight K included in both equations, commonly referred to as the K-factor, is the maximum possible adjustment per match. This weight is built into the Elo scoring system to serve as a linear adjustment proportional to the amount by which a player over-performs or under-performs their expected score. Therefore, when a player's actual score exceeds their expected score, the scoring system takes this as evidence that the player's rating is too low and needs adjustment upwards. Similarly, when a player's actual score falls short of their expected score, the scoring system takes this as evidence that the player's rating is too high and needs adjustment downwards. There is debate in the extant literature concerning the most accurate K-factor, yet it is typically set at $K = 16$ for master chess players and $K = 32$ for weaker players [5].

Sentiment Analysis

Natural language processing (NLP) is a popular branch of artificial intelligence concerned with allowing computers to understand text and spoken words in much the same manner as human beings do [6]. A major component of NLP is sentiment analysis, in which the focus is to properly identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions towards the subject. Sentiment that appears in text takes on two categories: *explicit*, where the subjective sentence directly expresses an opinion (e.g.,"It's a beautiful day"), and *implicit*, where the sentence implies an opinion (e.g., "The headphones broke in two days"). As expected, the former category tends to be much easier to analyze than the latter; therefore, it has been the main focus of research within this space thus far [7].

Due to its complexity, sentiment analysis is often broken down into three separate tasks: sentiment detection, polarity classification, and target identification. Sentiment detection involves the classification of text as objective or subjective. This classification is typically based on the examination of the adjectives included in sentences. Once the subjective text is properly identified, polarity classification is the next step. Given an opinionated piece of text, the goal is to classify the opinion as one of two opposing sentiment polarities (e.g., positive or negative) or locate its position on the spectrum between the two polarities. Note that the two steps defined above can be done at several levels (i.e., term, phrase, sentence, or document level). It is common practice to use the output of one level as the input for a higher level. The third and final step is target identification, which involves determining who expressed the discovered sentiment. This task is not very difficult when analyzing product or movie reviews; however; it can become complicated when considering forms of general writing such as webpages and blogs [7].

The applications of sentiment analysis in industry are plentiful. One major use case is found in call centers and chatbots, specifically the sentiment analysis of customer communications and call transcripts. More companies are using NLP to measure the sentiment expressed by customers about their products online and over the phone. If a company uncovers particularly negative sentiment surrounding a certain product, they can then devote more resources towards determining its major issues and resolving them. A very similar approach is being applied to measure employee satisfaction as well. Companies can collect transcripts of employee phone and chat conversations with their HR department, and use them to identify areas in which employees are struggling or not feeling supported. Once these struggle areas are identified, the company can then take the appropriate steps to improve the overall employee experience [8].

Elo Scoring & Sentiment Analysis at West Point

Given that the US Army recently awarded a contract to Aptima Inc. to overhaul soldier evaluation systems, and designated 2020 CST at West Point to serve as a trial run for newly digitized soldier tactical evaluation, researchers in the Systems Engineering Department pounced on the opportunity to change the archaic peer evaluation system. Mr. Ian Kloo, alongside fellow instructors Colonel Matthew Dabkowski and Colonel Riley Post, approached the leadership team within the Department of Military Instruction, whom is responsible for planning and executing CST, with a new and more quantitatively rigorous method to conduct peer evaluations. More specifically, they pitched the implementation of a comprehensive system that uses Elo scoring and sentiment analysis to create a virtual peer-ranking dashboard for each cadet.

Although Elo scoring is not typically used as a method for peer evaluations, it can be adapted very easily to suit this purpose. Consider a team consisting of n members. This results in $\frac{n(n-1)}{2}$ pairwise comparisons between team members. Each pairwise comparison is analogous to a match in chess. Every member of the team must select a "winner" and "loser" for each pairwise comparison (i.e., there are no "draws" between team members). Note that Elo scoring was designed to model scores for players over time; therefore, the order in which match-ups are evaluated impacts the score. In order to account for match-up order, the Systems Department proposed running a short simulation by shuffling match-ups and calculating the Elo scores for each squad member over many iterations. This results in an Elo score distribution for each cadet, which in turn can lead to a robust set of statistics that summarize each individual's performance in terms of peer evaluations. For instance, the Systems Department recommended using each cadet's distribution of Elo scores to calculate an average Elo score and create an appropriate ordering of all the members of a given team.

Additionally, to quickly summarize the qualitative assessments of squad mates that are provided by each cadet, the Systems Department proposed using sentiment analysis to classify the collected responses as favorable or unfavorable. This process would enable the senior Army officers charged with evaluating cadets to more easily identify the positive and negative comments tied to each individual. Furthermore, it would enable cadets, whom are often pressed for time at the conclusion of field training exercises, to sort through their peers' comments in a more efficient manner and gain valuable constrictive criticism.[1]

After the initial pitch, the Department of Military Instruction was immediately on board and eager to implement the Systems Department's solution to improving the peer evaluation system for summer training. Mr. Kloo and company got to work on the system used to gather data and build the dashboard for each cadet. He and his team successfully designed a data pipeline and interface through which cadets could complete pairwise comparisons and leave qualitative evaluations for each member of their squad. The system automated the process of data collection by sending an email to each cadets' standardized school email (first.last@westpoint.edu) with a link to the evaluation dashboard. Most cadets accessed the link using their personal devices; however, cadets without internet access were provided systems to complete their peer evaluations. Once at the evaluation dashboard, cadets picked between the pairwise comparisons of each squad mate presented to them and left comments for squad mates as they saw

[1] Refer to the attached html document or access our Deepnote notebook for an example of Elo scoring & sentiment analysis carried out in R.

fit. Cadet response data was then automatically transferred to the visualization dashboard which could be used for future counseling sessions and leadership development (see Appendix B for an example of a cadet's virtual dashboard).

After two summers of implementing the new peer evaluation system, the Department of Military Instruction noticed that they were gathering more robust, processable, and interpretable data from cadets. Additionally, cadets reported that they preferred to pick between two-peers rather than rank their entire squad because it made comparisons amongst individuals easier. In light of the success of the system prototype implemented by the Systems Department, the Department of Military Instruction has since moved their technology to the the William E. Simon Center for the Professional Military Ethic at West Point, where full time data analysts and software engineers continue to improve the system and draw leadership insights from the data collected on cadets.

Use of Peer Evaluations in Industry

While we discussed why peer evaluations are important in the public sector, specifically in the US Army, it is important to consider performance management and peer evaluations within the private sector as well. Two companies that have placed noticeable emphasis on these topics are Google and Netflix.

Few companies are as committed to performance management as Google. Over the years, they have invested millions in developing its people operations systems and practices. In fact, Google's SVP of People Operations, Laszlo Bock, authored *Work Rules* detailing Google's innovation, experiments, and R&D in establishing its HR functions. Google has instituted a detailed performance review system to retain its very best talent (smart creatives as they like to refer to their employees) beyond financial incentives and perks.

In order to assess its employees, Google conducts a two-phased annual performance review. First, a preview is carried out at the end of the first semester, followed by a complete review between October and November. The performance evaluation constitutes a self-assessment and 360-degree feedback (Google's peer evaluation system). Google incorporated the 360-degree feedback to eliminate potential bias in self-assessment as well as to provide a more holistic picture of employees' performance. As part of its approach, managers ask employees to create a shortlist of peers who can testify to the accuracy and authenticity of the self-assessment. The peers, across functions, assess the employee on their contribution, impact, strengths, and weaknesses. This helps managers assess and verify the self-assessment report [9, 10].

Netflix recently followed suit, putting an end to its formal performance reviews and instead instituting informal 360-degree reviews. However, unlike Google, Netflix has kept its review process fairly straightforward. It asks its employees to identify things that their colleagues should stop, start, or continue. While Netflix originally used an anonymous software system, eventually they moved to using identified feedback. Also, unlike the twice a year approach employed by Google, Netflix's reviews are much more regular and need-based

[11].

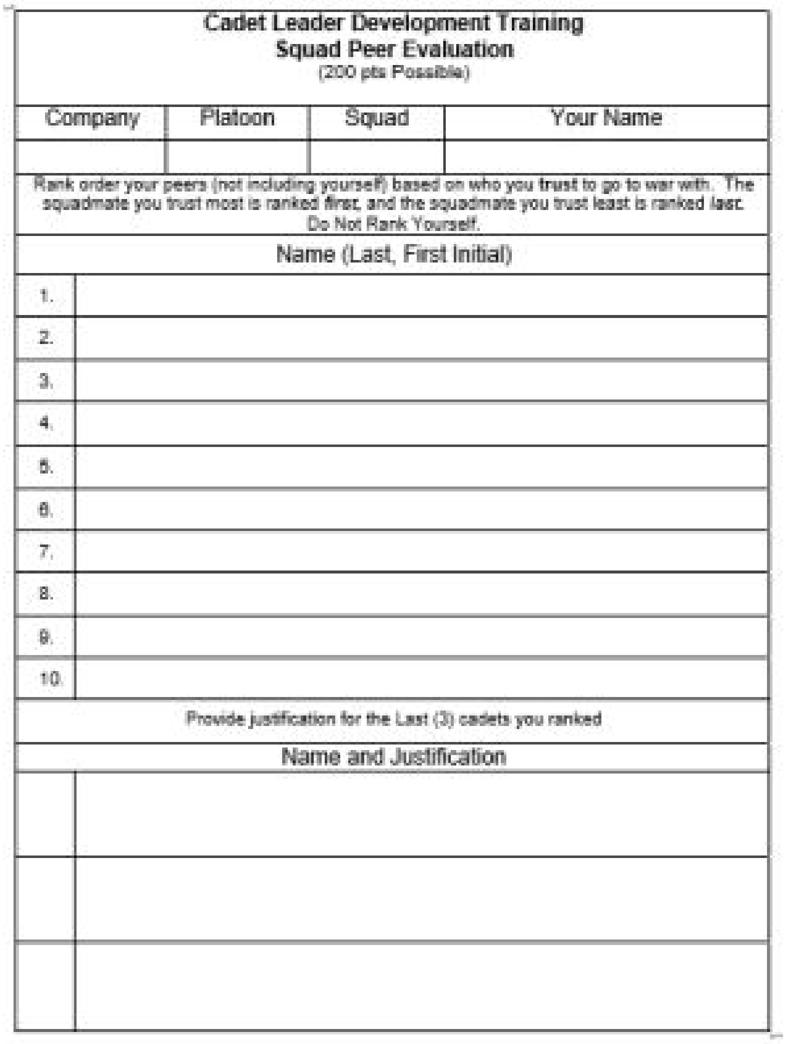
Perhaps it would be intriguing to pilot an Elo scoring system for hard-skilled roles at Google and Netflix. Since the performance reviews for these roles skew heavily towards achieving specific results, an Elo-scoring approach could potentially help to more clearly rank employees according to their functions and skills. Further, at Google, this ranking could potentially be integrated into the Chameleon program to optimize the staffing and business needs. Evaluating the results of an Elo-based approach with the 360 degree feedback system that Google and Netflix currently employ could yield additional benefits. Additionally, because Elo scoring was designed to model performance overtime, employers could use it to develop a continuous performance feedback system as opposed to the current practice of providing feedback twice a year. Regardless, whether or not the private sector chooses to adopt Elo scoring, it could benefit from using more rigorous quantitative methods associated with peer evaluation.

Typically, public sector tech initiatives mimic successful private sector implementations. However, due to the synthesis of ideas in the Systems Department and unique training environment at West Point, using Elo simulations and sentiment analysis in peer evaluations seems like an effective, novel approach. While we propose many avenues for private-sector adoption, considering how such a system would fit into the culture of an organization is important as well. Is the Elo method a 100% approved solution for peer evaluations? Probably not, but it could yield new insights for high performing corporate teams.

Acknowledgements: We would like to thank Mr. Ian Kloo, West Point Systems Engineering Department, and Major Casey Williams, West Point Department of Military Instruction, for meeting with us, providing us with context, implementation details, and the materials included in the appendices.

Appendix A: Traditional Peer Ranking System

Figure 1 included below is an example of the traditional peer ranking card that cadets were required to fill out at the conclusion of each summer training exercise prior to 2019.



Cadet Leader Development Training
Squad Peer Evaluation
(200 pts Possible)

Company	Platoon	Squad	Your Name

Rank order your peers (not including yourself) based on who you trust to go to war with. The squadmate you trust most is ranked first, and the squadmate you trust least is ranked last. Do Not Rank Yourself.

Name (Last, First Initial)

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

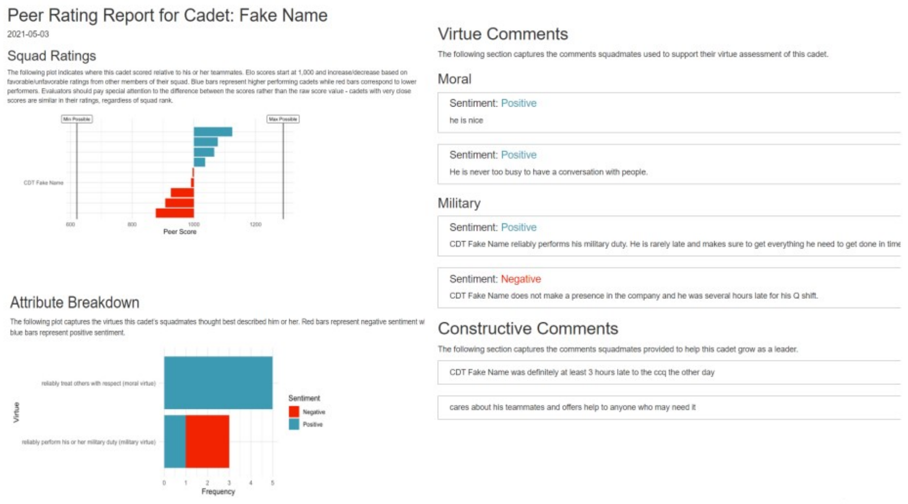
Provide justification for the Last (3) cadets you ranked

Name and Justification

Figure 1: Traditional Peer Ranking Card

Appendix B: Improved Peer Ranking System

Figure 2 included below is an example of the new and improved peer ranking dashboard that is currently being used. Note the squad ratings included on the left and the sentiment analysis included on the right.



Peer Rating Report for Cadet: Fake Name

2021-05-03

Squad Ratings

The following plot indicates where this cadet scored relative to his or her teammates. Elo scores start at 1,000 and increase/decrease based on favorable/unfavorable ratings from other members of their squad. Blue bars represent higher performing cadets while red bars correspond to lower performers. Evaluators should pay special attention to the difference between the scores rather than the raw score value - cadets with very close scores are similar in their ratings, regardless of squad rank.

Attribute Breakdown

The following plot captures the virtues this cadet's squadmates thought best described him or her. Red bars represent negative sentiment w
blue bars represent positive sentiment.

Virtue
reliably treat others with respect (moral virtue)
reliably perform his or her military duty (military virtue)
Sentiment
🟥 Negative
🟦 Positive
0
1
2
3
4
5
Frequency

Virtue Comments

The following section captures the comments squadmates used to support their virtue assessment of this cadet.

Moral

Sentiment: Positive
he is nice

Sentiment: Positive
He is never too busy to have a conversation with people.

Military

Sentiment: Positive
CDT Fake Name reliably performs his military duty. He is rarely late and makes sure to get everything he need to get done in time

Sentiment: Negative
CDT Fake Name does not make a presence in the company and he was several hours late for his Q shift.

Constructive Comments

The following section captures the comments squadmates provided to help this cadet grow as a leader.

CDT Fake Name was definitely at least 3 hours late to the ccq the other day

cares about his teammates and offers help to anyone who may need it

Figure 2: Improved Peer Ranking Dashboard

References

[1] Arpad E Elo. “New USCF Rating System”. In: *Chess Life* 16 (1961), pp. 160–161.

[2] Arpad E Elo. *The Rating of Chessplayers, Past and Present.* BT Batsford Limited, 1978.

[3] Raghav Mittal. “What is an Elo Rating?” In: *Purple Theory* (2020).

[4] Paul CH Albers and Han de Vries. *Elo-Rating as a Tool in the Sequential Estimation of Dominance Strengths.* 2001.

[5] Mark E Glickman and Albyn C Jones. “Rating the Chess Rating System”. In: *CHANCE-BERLIN THEN NEW YORK-* 12 (1999), pp. 21–28.

[6] IBM Cloud Education. “Natural Language Processing (NLP)”. In: *IBM* (2020).

[7] Yelena Mejova. “Sentiment Analysis: An Overview”. In: *University of Iowa, Computer Science Department* (2009).

[8] Maria Korolov. “What is Sentiment Analysis? Using NLP and ML to Extract Meaning”. In: *CIO* (2021).

[9] Lori Li. “How Google evolved performance management to drive top performance across its growing workforce”. In: *TinyPulse* (2022). URL: `https://www.tinypulse.com/blog/how-google-evolved-performance-management-to-drive-top-performance-across-its-growing-workforce`.

[10] Shana Lebowitz. “How Google Manages Performance Reviews for Employees”. In: *Business Insider* (2021). URL: `https://www.businessinsider.com/how-google-performance-reviews-work-2015-6`.

[11] Ashley Rodriguez. “Netflix’s culture of intense feedback comes to a head during spring review season. Insiders describe how it works.” In: *Business Insider* (2020). URL: `https://www.businessinsider.com/how-google-performance-reviews-work-2015-6`.