

Describing Success in U.S. Army Officers Across Different Branches: An Analysis of Officer Evaluation Reports

Heidy Shi, JD Caddell, and Julia Lensing

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

Corresponding author: heidy.shi@westpoint.edu

Author Note: Cadet Heidy Shi is a senior at the United States Military Academy at West Point majoring in Systems Engineering with a minor in Applied Statistics. This research was completed as part of the requirements for the Honors Program under the advisement of JD Caddell and Julia Lensing. Cadet Shi would like to thank the Department of Systems Engineering and Army Human Resources Command for their support of this research.

Abstract: Each job field (branch) in the Army requires a unique set of skills and talents of the officers assigned. Officers who demonstrate the required skills are often more successful in their assigned branch. To better understand how success is described across branches, research was conducted using text mining and sentiment analysis of a data set of Officer Evaluation Reports (OERs). This research looked for common trends and discrepancies across varying branches and like groups of branches. The research also examined the sentiment conveyed in the narrative portion of OERs by analyzing words and bigrams commonly used to describe varying degrees of performance by officers. Findings show that qualitative narratives for the top two performance designations fail to differentiate between officers' varying levels of performance regardless of branch.

Keywords: Text Mining, Sentiment Analysis, Officer Evaluation Reports (OERs)

1. Introduction

1.1 Background on the Officer Evaluation System and Officer Evaluation Report

Like all enterprises, the U.S. Army is continually seeking the best method for evaluating performance of its officers. The current Officer Evaluation Report (OER) has been a product of decades of research and development and has a significant impact on the careers of all Army Officers. The OER provides necessary feedback to the officer on their performance of duties and provides senior officers the information required to make decisions regarding the officer's future career (Kite, 1998). These decisions include promotions, assignments, selections for advanced schooling, and retention on active duty (Straffon, 1997).

The newest changes to the officer evaluation system were implemented in 2013. These changes included establishing a distinction between the primary rater and senior rater, altering the performance designations, and implementing a new rater profile to keep track of all OERs completed. (Lopez, 2013). Currently, every officer receives a yearly evaluation with two ratings, one from their first-line supervisor (primary rater) and one from the next higher supervisor in the chain of command (senior rater). The primary rater's responsibility is to evaluate the officer's performance of duties, based on professionalism, competencies, and attributes. The senior rater's responsibility is to focus on the officer's potential for future service and additional responsibility through the ranks. The OER is structured into four blocks that each rater and senior rater can catalogue an officer's performance compared to others. The rater levels are "excels" which encompasses only the top 49%, "proficient", "capable", and "unsatisfactory." The senior rater levels are "most qualified" which also encompasses only the top 49%, "highly qualified", "qualified", and "not qualified." These block checks (performance designations) are often the first data point examined when evaluating potential to retain or promote officers by promotion board. These quantitative block checks are paired with a section for qualitative comments where raters and senior raters can input remarks (Department of the Army, 2015).

To avoid over inflation of the evaluation system and to delineate high performers amongst the ranks, the current OER system uses a forced distribution ranking structure by limiting the percentage of "most qualified"/ "excels" (top block check) that both raters senior raters may assign. Both raters and senior raters must maintain a "credible" rater profile by assigning less than 50% of the ratings in the top block for any given rank. As part of maintaining a credible

profile, senior raters are encouraged to “maintain a “cushion” in the number of “most qualified” ratings given to prevent exceeding the 50% threshold (Department of the Army, 2015). However, simply selecting a block check may not always tell the full story of officer’s performance. Ostensibly, the qualitative narrative portion of the report holds merit as well in describing performance and future potential as an officer in the Army. Generally, officers feel the senior rater’s narrative has more impact than the rater’s narrative and thus became the focus of this research. Through text analysis this research sought to identify patterns and trends within these narratives.

1.2 Motivation for Research

1.2.1 Limitations to Senior Rater Block Checks

Due to the limitation identified above and the requirement to maintain a credible profile, senior raters are cautious in the execution of their rankings. Officers with immature rating profiles may choose to reserve top block checks during their initial ratings to build flexibility for their profile. On the other hand, officers who mismanage their profile are often unable to reward officers they feel are deserving of a top block rating (Cho, 2015). Inevitably, the system may force the hand of a rater or senior rater to assign a rating that is not a true representation of an officer’s performance or potential and reduces some of the credibility that the block checks should provide.

Additionally, senior raters are not allowed to specifically say that they were unable to give out a top block check because of their senior rater profile limitations (Department of the Army, 2015). As a result, a common assumption is that senior raters often focus on the narrative portion of the OER to try and draw attention to an officer’s potential, despite not being able to give out a top block check. Senior raters often put a lot of effort into their senior rater narratives, sometimes choosing to focus on enumeration to try and make certain officers stand out. There may also be certain words that senior raters will use that have more meaning and a stronger effect to portray future potential amongst the officers they rate. This research will test these assumptions to see if there are any trends or discrepancies across OER narratives.

1.2.2 Talent Based Branching

Each branch requires specific talents of its officers and the jobs performed by each officer vary significantly across branches. Each branch publishes a talent “storyboard” that details specific native intelligences, skills, knowledge, and behaviors demanded by each of the 17 basic branches. There are some talents that are in high demand across several branches as well as some heterogeneity across the branches. Table 1 outlines the branch talent demand across branches and indicates that there are closer talent correlations among maneuver branches such as Infantry (IN) and Armor (AR), just as there are correlations among logistics and sustainment branches such as Quartermaster (QM), Ordnance (OD), and Transportation (TC) (Calarusso, Heckel, Lyle, & Skymmyhorn, 2016). Naturally, one would assume there should be discrepancies across the different branches regarding the words senior raters use to describe success and failure amongst Army officers. More interestingly, there may also be common trends across each of the branches that senior raters use to define success. This research will test these assumptions as well to see if there are any trends across different branches.

Table 1. Talent Requirement Matrix by Branch (Calarusso, Heckel, Lyle, & Skymmyhorn, 2016)

Talent Demand	AD	AG	AR	AV	CM	CY	EN	FA	FM	IN	MI	MP	MS	OD	EOD	QM	SC	TC
Bodily-Kinesthetic																		
Communicator	1	1	1		1				1	1	1		1	1	1	1		
Cross-Culturally Fluent	1								1			1						1
Detail-Focused		1											1		1			
Innovative				1	1	1	1			1								1
Inspirational Leader	1				1	1				1								
Interdisciplinary				1				1			1	1	1					
Interpersonal		1		1								1	1	1				1
Logical/Analytical	1	1			1	1	1		1		1			1				1
Mentally Tough			1		1			1		1		1			1	1	1	1
Multi-Tasker	1	1	1	1				1		1		1				1	1	1
Perceptive											1	1					1	
Physically Fit	1		1		1			1		1		1						
Problem Solver			1			1	1			1	1			1	1		1	1
Process Disciplined								1	1							1		
Project Manager								1	1					1			1	1
Prudent Risk Taker			1	1												1		
Spatially Intelligent				1				1										
Technologically Adept						1												1

Note: AD = Air Defense,
 AG = Adjutant General,
 AR = Armor, AV = Aviation,
 CM = Chemical Corps,
 CY = Cyber, EN = Engineers,
 FA = Field Artillery, FM = Finance,
 IN = Infantry, MI = Military
 Intelligence, MP = Military Police,
 MS = Medical Service Corps,
 OD = Ordnance, EOD = Explosive
 Ordnance Disposal,
 QM = Quartermaster, SC = Signal
 Corps, TC = Transportation

1.3 Officer Evaluation Report Data Set

1.3.1 Make-Up of the Data

The data set used to conduct this analysis includes every active duty OER that was submitted with an end date during the 2017 evaluation period. The data set includes over 156,000 entries and is organized based on six factors: gender, branch, rater performance designation (block check), rater narrative, senior rater performance designation (block check), and senior rater narrative. The gender was coerced based off pronouns used within the narrative of each entry. The branches in the data set include all functional areas and 17 basic branches. All data was de-identified to protect confidentiality of the rates and raters. The rank of the rated officer was also redacted, but the rank for the data set ranged from warrant officers to colonels. Figure 1 is a summary of the total number of OERS per branch broken down by the senior block check rating. This data reflects the relative size and density of officers within each branch. For example, while Aviation (AV) is not the largest branch in the Army, it has a high density of both warrant and commissioned officers as reflected in their 9.2% of total evaluations in 2017.

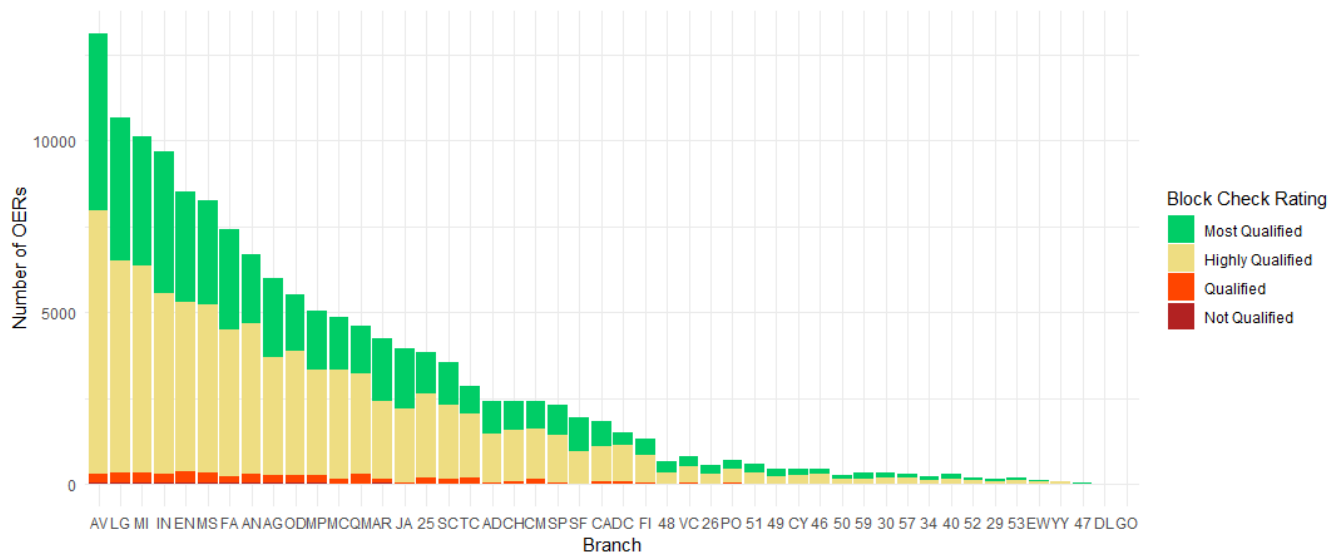


Figure 1. Total Number of OERS in 2017 by Branch

1.3.2 Limitations to the Data

The narrative portion of the OER is similar to an essay evaluation, which is one of the six categories of methods of performance appraisal often utilized by the U.S. business and industrial sector (Morrisey, 1983). While essay evaluations are excellent for capturing details and providing specific internal feedback, the greatest disadvantage is that they are subjective in nature and much more difficult to use when comparing others. Additionally, essay evaluations are limited to the rater’s ability to provide good feedback. Often times, ratees who are evaluated by raters who are better at articulating and writing evaluations may come off as stronger performers than those being rated by raters with a poor writing ability (Milkovich, 1997). Consequently, focusing on specific words may not be able to accurately tell the full story of how the Army describes officer success.

Additionally, the block check portion of the OER is based on a forced distribution rating system to combat inflated ratings and to clearly distinguish high-performing officers from their peers. Proponents of forced distribution rating systems believe that the system motivates individuals, eliminates “dead wood,” forces raters to be honest with their ratees, and develops strong leaders. It also benefits poor performers by encouraging them to move on to other jobs for which they may be better suited (Welch, 2001). On the other hand, forced distribution systems can also be viewed as unfair, subjective, and vulnerable to biases. It also may discourage collaboration and teamwork (Pfeffer & Sutton, 2000). Thus, looking at the forced distribution of block checks also may not tell the full story of officer success.

Moreover, the current evaluation system limits the data set to provide a perspective of only two superior officers. As a result, these measures only look at mission accomplishment and do not measure molding and motivating soldiers and units

for long-term success (Reese, 2002). There may also be other factors that breed officer success that are not described in the OER. The data set also limits the ability to test the true impact of these evaluations because there is no way to tell whether the officers being rated were actually promoted or selected for key positions.

2. Methodology

2.1. Exploratory Data Analysis and Data Mining

Exploratory data analysis is defined as the process of discovering interesting patterns and knowledge from large amounts of data. Steps that occur during this process include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. During data cleaning, inconsistent data and noisy data are removed because it may detract from the data mining process and pattern evaluation. Data transformation allows data to be transformed and consolidated into forms appropriate to perform both summary and aggregate operations. Then, through data mining, patterns are extracted and evaluated to produce results and conclusions (Han & Pei, 2011).

Data mining was first conducted on the given data set to ensure the data was usable to produce patterns. The data was cleaned to eliminate all O6 (Colonel) OER narratives, which were filtered based on specific block checks used only in the evaluation of colonels. Empty data fields where no senior narrative was present were also eliminated. Once the data was cleaned, exploratory data analysis was conducted on the new data set to see if it was an accurate representation of the true OER profile. It also gives a better understanding of the overall data set and allows for discovery of discrepancies to focus on when conducting further research. Figure 2 shows that the overall distribution of OERs given does provide an accurate representation of how block checks should be specified. There were less than the 50% allotted for “most qualified” block checks, with only 37% of all OERs in the 2017 rating period being given a “most qualified” rating. Most of the OERs were given a “highly qualified” block check, as almost 60% of the OERs were rated “highly qualified.” There were only a small portion of “qualified” and “not qualified” ratings given out. Oftentimes, these ratings are only given out for extremely poor performance or disciplinary issues.

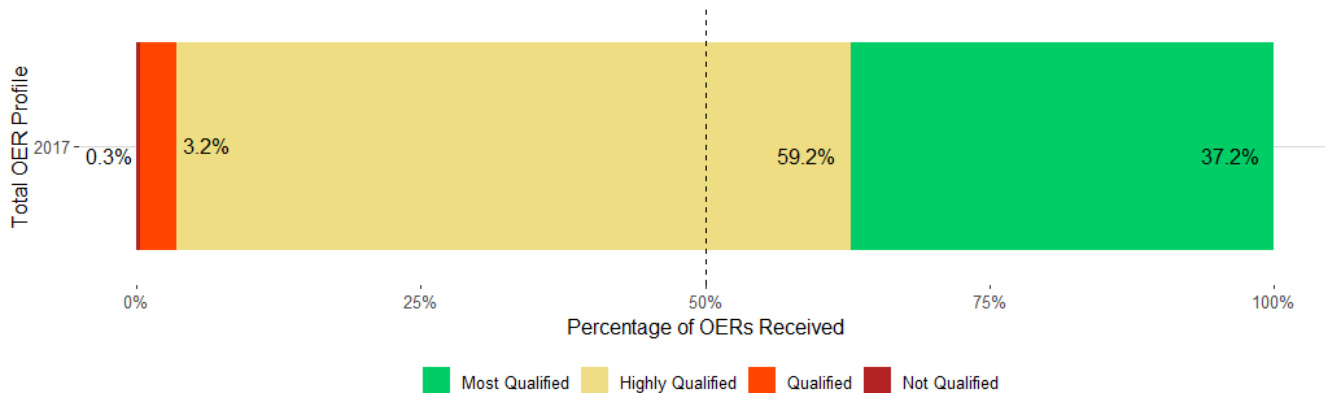


Figure 2. Distribution of OERs by Senior Rater Label Block Check During the 2017 Evaluation Period

After understanding the overall distribution of block checks across all OERs given, it is important to compare that distribution across the branches to see if block checks across the branches were similar to block checks across the whole Army. Figure 3 shows that across all basic branches in the given data set that had more than 100 OER evaluations in 2017, the percentage of top blocks given by the senior rater were distributed relatively the same as that of the whole Army. Additionally, none of the branches exceeded in giving out more than the allotted 50% of “most qualified” block checks. Each branch stayed well within the 50% threshold, with most OERs given as “highly qualified”. This information can suggest that senior raters are maintaining a credible profile regardless of the branch they are in. Moreover, it suggests that officers across the branches are also receiving an equal rating because one branch does not give out more “most qualified” block checks than any other branch.

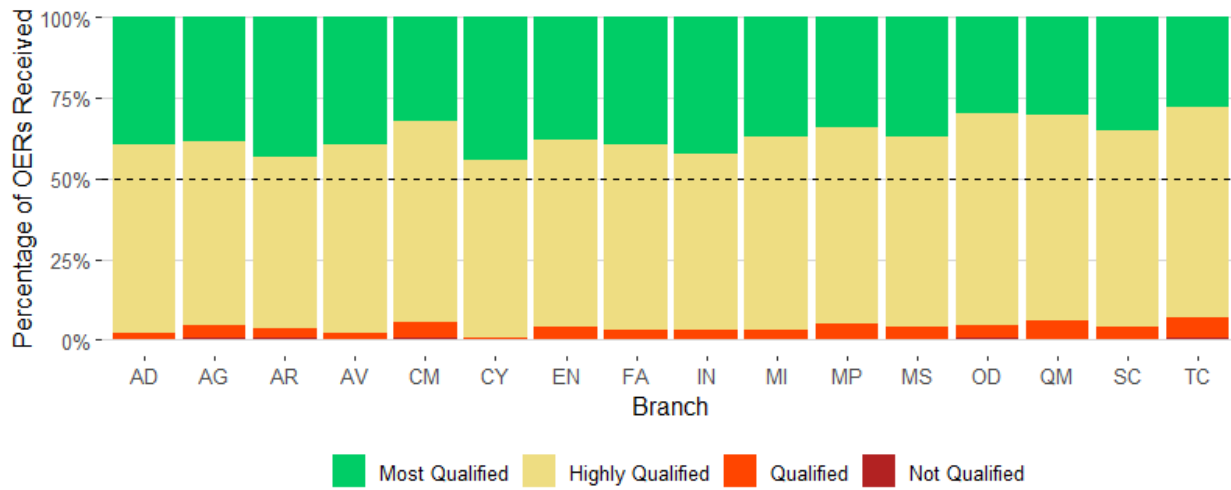


Figure 3. Distribution of OERs Across Base Branches in 2017

2.2 Text Mining

Text mining discovers and analyzes information, specifically text within documents, to discover patterns. Text mining goes beyond information access to help users analyze and digest information and facilitate decision making. The main goal of text mining is to look for trends and outliers amongst text data. (Aggarwal & Zhai, 2012).

Part of text mining that researchers are concerned with is text frequency (*tf*), which is how often certain words appear within a document of text. Another key component of text mining is inverse document frequency (*idf*) which decreases the weight for commonly used words and increases the weight for words that are not used as frequently in a collection of documents. The *idf* for a given term can be calculated using the equation

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right) \quad (1)$$

Inverse document frequency is often combined with term frequency to calculate a term's *tf-idf* (the two quantities multiplied together) to find the frequency of a term adjusted for how rarely it is used. The statistic *tf-idf* is intended to measure how important a word is to a document in a collection of documents (Silge & Robinson, 2019). To understand word frequencies in OERs, the *tf-idf* will look at what words senior raters are using in their senior rater comments using the tidytext package in R. Looking at the *tf-idf* in senior rater comments will help show what words are frequently used but are more important for specific block checks and specific branches. It can help pinpoint what words are frequently used but also distinct to each branch.

Another technique commonly used in text mining is to look at the frequency of consecutive words. Looking for patterns in consecutive words allows researchers to analyze relationships between words. Some common relationships include examining which words tend to follow others immediately or which words tend to occur together across the same documents. Looking at pairs of words also provides more insight to the context of common words used within documents. In order to analyze consecutive words, “n-grams” are used to tokenize adjacent pairs, where “n” is the number of adjacent words being paired together. When looking at consecutive words, researchers will commonly study pairs, also known as bigrams, or triples, also known as “trigrams” to find relationships and correlations between multiple words (Silge & Robinson, 2019).

3. Results

3.1 Text Mining for Trends Within Branches

Text mining was conducted on the OER data to look at *tf-idf* across senior rater comments. Figure 4 expresses the results of text mining senior rater comments across the maneuver branches (Armor, Infantry, and Field Artillery) and the logistics branches (Ordinance, Quartermaster, and Transportation). The top 15 words with the greatest *tf-idf* are shown for

each branch. Naturally, each branch has specific words that senior raters use to talk about the officers they rate. For example, common words used by Armor officers include “reconnaissance” and “scout”, whereas words used to describe Infantry officers include “maneuver,” “mortar,” and “rifle.”

When looking at the logistics branches, common words used amongst the three branches include “logistics.” There are also discrepancies across the branches. The Ordnance branch focuses a lot of its time on maintenance, hence “maintenance,” and “ammunition” were some of the top words found in senior rater narratives for that branch. Quartermaster focuses on other aspects of supply, hence “food”, “water” and “petroleum” are some of the top words. Transportation also yielded special words across its narratives to include “watercraft” and “vessel” which accurately describes the roles Transportation officers have. The *tf-idf* for each word sorted by the branches helps identify the words that are important to each of the branches amongst the collection of OERs. It shows that these words are what distinguishes the OER narratives apart across each of the different branches.

The words identified show that the words senior raters are using are reflective of the jobs the rated officers have held. These words do not describe attributes that they have demonstrated for their particular branch. Specific adjectives identifying talent and potential also do not appear. Additionally, since *tf-idf* only identifies unique words specific to each branch, the pattern suggests that specific adjectives and descriptors that are used may be so common across all branches that it is not reflected at all when analyzing OERs using *tf-idf*. As a result, the only words used by senior raters to differentiate narratives across the branches are words indicative of past jobs held or possible future job opportunities.

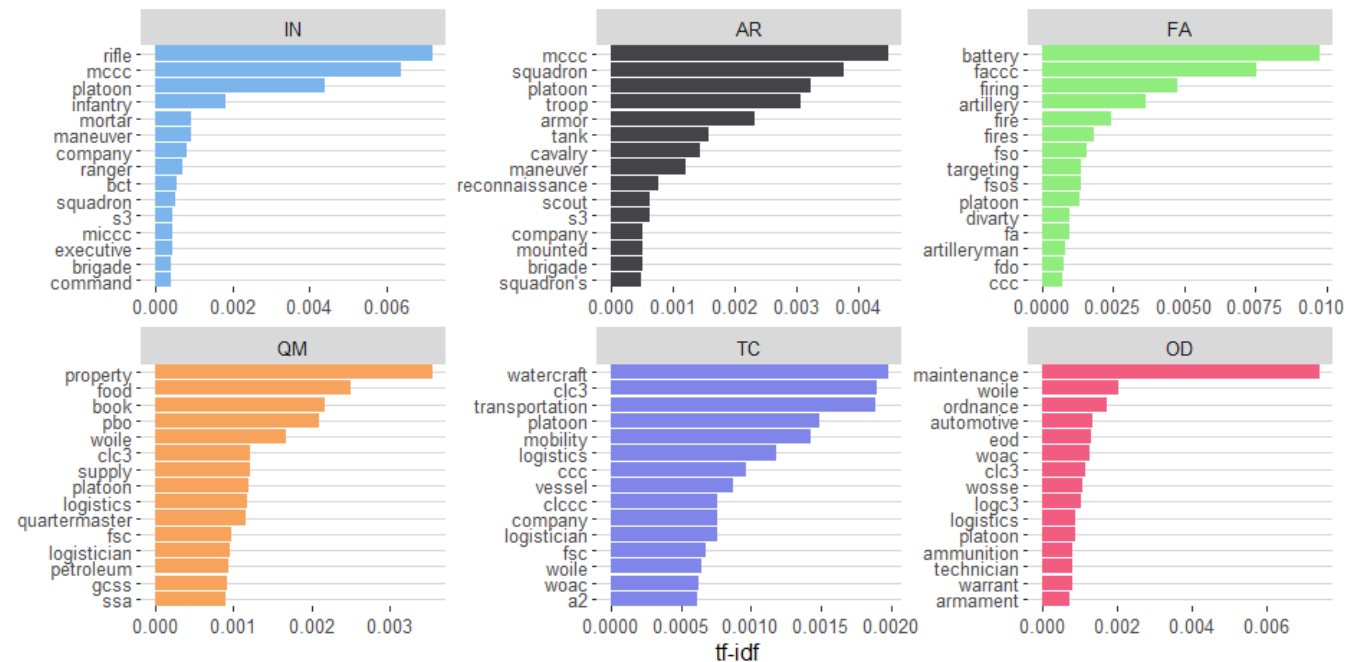


Figure 4. Text Mining Across Maneuver and Logistics Branches

3.2 Text Mining by Bigrams

Text mining using bigrams was also conducted to look at patterns of consecutive words used in OERs. Figure 5 displays four graphs of the most frequent bigrams across the maneuver branches and the logistics branches, filtered by two block checks, “most qualified” and “highly qualified.” By analyzing these bigrams across the “most qualified” and “highly qualified” narratives, patterns can be discovered to see if the same words by senior raters are being used, regardless of the block check given.

Across the maneuver branches, one of the most common bigrams used to talk about the “most qualified” officers is “unlimited potential.” This bigram seems to be the key word that most senior raters use as the frequency of the bigram is significantly higher compared to other bigrams. Senior raters are also choosing to use enumeration by using words such as “top 5,” and “top 10” to talk about their top performing officers within their branch. Additionally, senior raters are frequently

writing “promote ahead” which is consistent with the meaning of giving out “most qualified” block checks. Looking at the bigrams used in “highly qualified” ratings, it is interesting to note that the common words used include “unlimited potential” and “promote ahead.” When comparing the words used in “most qualified” and “highly qualified” narratives, the words used are very similar, which may indicate that the narratives between the block checks given are very similar, regardless of the block check given. Ultimately, this could result in two different possibilities. One, officers within the maneuver branches may not be distinguished from one another by the narratives given, and only by the block checks provided. Second, officers within the maneuver branches are using the same words in the narratives given because they are unable to give out the top block check and are attempting to compensate with the narrative. Either way, both cases indicate that there is a decrease in the relative importance of the narratives. As currently written, the narrative serves as an attempt to circumvent the delineation process and makes the OER a less effective evaluation tool for the Army.

Across the logistics branches, the most common bigrams used to talk about the “most qualified” officers is also “unlimited potential” and “promote ahead” with enumeration words such as “top 5” and “top 10” as the next most frequent words used. “Promote ahead” and “promote immediately” are also common words, which confirms that the words senior raters are using to talk about their “most qualified” officers is consistent with the intent of giving out the top block check. When looking at the bigrams used in “highly qualified” ratings, it is interesting to note that once again, the same words are being used in “most qualified” and “highly qualified” OER narratives. Similar to the maneuver branches, the logistics branches reveal the same pattern between “most qualified” and “highly qualified” senior rater narratives. Furthermore, the words used in both the logistics and maneuver branches are almost identical. This pattern may indicate that there is a common standard that senior raters use when constructing their top block narratives.

Most interestingly, when analyzing enumeration words across “most qualified” and “highly qualified” officers, regardless of the branch, the same words are also being used. Senior raters are frequently writing “top 5” or “top 10” in both “most qualified” and “highly qualified” narratives, which could indicate some disparity. This is because officers that are in the top half of the population should all be receiving “most qualified” block checks. Thus, those receiving a “highly qualified” block check but still have “top 5” or “top 10” written in their narrative suggests that the block check and the narrative do not match. The narrative may be trying to portray that the block check is not reflective of the officer’s true potential.

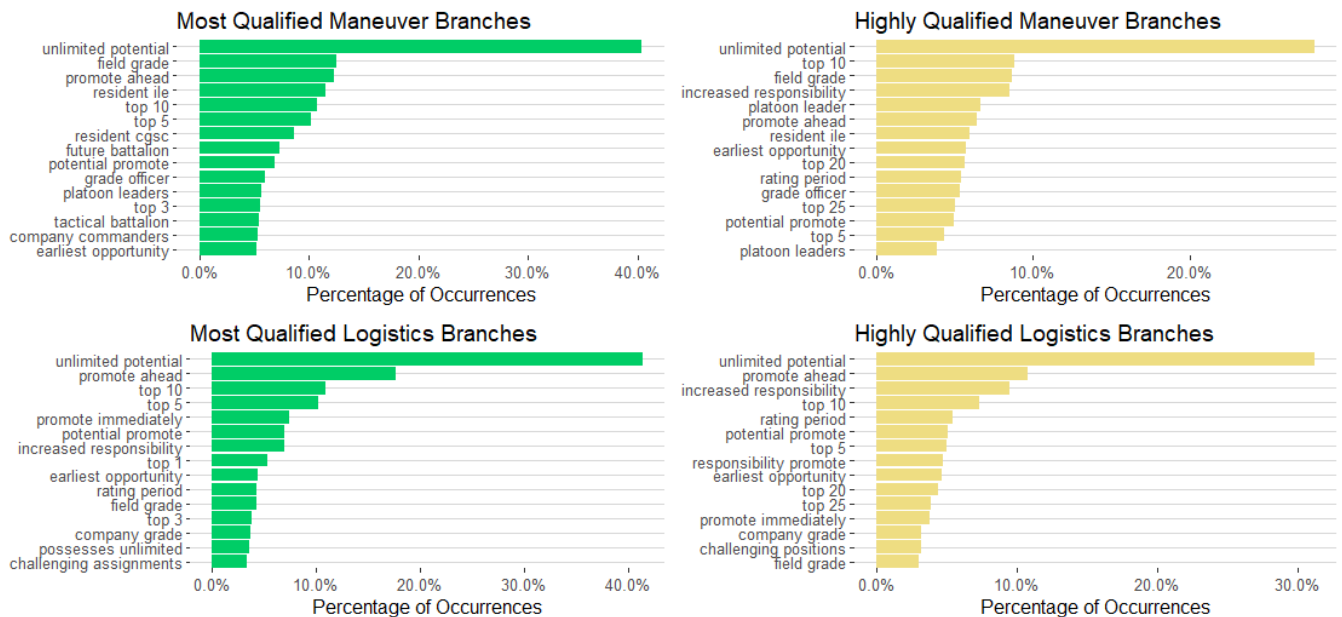


Figure 5. Frequency of Bigrams in “Most Qualified” and “Qualified” OERS Across Branch Groups

4. Conclusion and Further Research

The OER and the officer evaluation system poses many limitations on senior rater ability to accurately reflect an officer's performance and potential. These limitations can be significant particularly when evaluating for promotions, assignments, selections for advanced schooling, and retention on active duty. The rater narratives have become exceedingly more important to paint a picture of what the rated officer has accomplished during that evaluation period. Across the branches, senior raters are using very specific words relevant to their own branches to talk about the successes and failures of other officers. However, these specific words only focus on jobs previously held or recommend future jobs that officers may hold. Descriptive words focusing on talent and skills that should be different for each branch are not present and do not align with the Army's intent of stressing talent-based branching. More interestingly, officers use relatively the same words to talk about their "most qualified" and "highly qualified" officers across maneuver and logistics branches. This could indicate that officers across the branches have adapted a common standard to talk about officers in their narratives, irrespective of the block check or the branch. This can create many problems because it shows that the narrative portion does not make any distinctions between its officers. Key words in a "highly qualified" narrative can look exactly like key words in a "most qualified" narrative. As a result, the narrative portion may not be valued as much as the block check and may be obsolete in evaluating officer performance and potential.

The 2017 OER data set provides many opportunities to conduct further research to gain a better understanding of how raters speak about officer performance and potential. Further research can compare the narratives of primary raters to senior raters to see if there are any patterns or discrepancies between what primary raters and senior raters are saying about their top-rated officers. The Army also provides other guidance regarding what should be included in rater narratives and encourages raters to use enumeration when writing their narrative. Further research could investigate how enumeration affects block checks and whether or not raters are being consistent with the Army guidance given on how to write OERs.

Understanding how officers in the Army are being evaluated is important and can provide new insights into changes to regulations regarding how OERs need to be conducted. If discrepancies are found, they can lead to new changes to continue to improve the overall officer evaluation system so that officers in the Army are receiving evaluations that accurately reflect their performance and potential and serve as a tool to identify the highest performers.

5. References

- Aggarwal, C.C., & Zhai, C. (2012). *Mining Text Data*. New York, NY: Springer Science + Business Media, LLC.
- Calarusso, M.J, Heckel, K.G., Lyle, D.S., & Skimmyhorn, W.L. (2016). Starting Strong: Talent-Based Branching of Newly Commissioned U.S. Army Officers. *Strategic Studies Institute and U.S. Army War College Press*, 9(1), 1-62.
- Department of the Army. (2015). *Evaluation Reporting System*. (Army Regulation 623-3). Washington DC: U.S. Government Printing Office.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. Waltham, MA: Elsevier Inc.
- Kite, D.P. (1998). *The US Army Officer Evaluation Report: Why are We Writing to Someone Who Isn't Reading?* Air Command and Staff College Air University.
- Lopez, T.C. (2013, March 29). "New Army OER Means Fewer Boxes, More Accountability for Raters." *Army News Service*. Retrieved from <https://www.jble.af.mil/News/Article-Display/Article/257909/new-army-oes-means-fewer-boxes-more-accountability-for-raters/>
- Milkovich, G.T., & Boudreau, J.W. (1997). *Human Resource Management*. Boston, MA: Times Mirror Higher Education Group.
- Morrissey, G.L (1983). *Performance Appraisals for Business & Industry*. Reading, MA: Addison-Wesley Publishing Company.
- Pfeffer J, Sutton RI. (2000). *The knowing-doing gap: How smart companies turn knowledge into action*. Boston: Harvard Business School Press.
- Reese, T.R. (2002). *Transforming the Officer Evaluation System: Using a 360 Degree Feedback Model*. US Army War College, Pennsylvania.
- Silge, J., & Robinson, D. (2019). *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O'Reilly Media, Inc.
- Straffon, N. (1997). "Promotion Boards." *Army Reserve Magazine* (Winter), 19.
- Welch, J.F. (2001). *Jack: Straight from the gut*. New York: Warner Books.