# Simulation-based analysis of a forced distribution performance appraisal system

Lee Evans and Ki-Hwan Bae
*Department of Industrial Engineering, University of Louisville, Louisville, Kentucky, USA*

Received 21 October 2017
Revised 15 December 2017
Accepted 18 December 2017

## Abstract

**Purpose** – The paper aims to estimates the limitations of a forced distribution performance appraisal system in identifying the highest performing individuals within an organization. Traditionally, manpower modeling allows organizations to develop plans that meet future human resource requirements by modeling the flow of personnel within an organization. The aim is to quantify the limitations of a performance appraisal system in identifying the best-qualified individuals to fill future requirements.

**Design/methodology/approach** – This paper describes an exploratory study using discrete event simulation based on the assignment, evaluation and promotion history of over 2,500 officers in the US Army. The obtained data provide a basis for estimating simulation inputs that include system structure, system dynamics, human behavior and policy constraints. The simulation approach facilitates modeling officers who receive evaluations as they move throughout the system over time.

**Findings** – The paper provides insights into the effect of system structure and system dynamics on the evaluation outcome of employees. It suggests that decreasing the number of a rater's subordinates has a significant effect on the accuracy of performance appraisals. However, increasing the amount of time individuals spend on each assignment has little effect on system accuracy.

**Practical implications** – This research allows an organization's leadership to evaluate the possible consequences associated with evaluation policy prior to policy implementation.

**Originality/value** – This work advances a framework in assessing the effect of system dynamics and structure, and the extent to which they limit or enhance the accuracy of an organization's forced distribution performance appraisal system.

**Keywords** Military, Performance appraisal, Discrete event simulation, Forced distribution, Manpower modeling

**Paper type** Research paper

## Introduction

Accurately identifying high performing officers is a key component of the Army Leader Development Strategy (Army Leader Development Strategy, 2013). Due to the Army not routinely recruiting mid and senior-level leaders, they must identify high performers and assign them to a broad range of opportunities to develop senior leaders prior to the need for the individual. Traditionally, manpower modeling allows organizations to develop plans that meet future human resource requirements by modeling the flow of personnel within an

Journal of Defense Analytics and Logistics
Vol. 1 No. 2, 2017
pp. 120-136
Emerald Publishing Limited
2399-6439
DOI 10.1108/JDAL-10-2017-0022

organization. In this work, we provide a framework assessing the effect of system dynamics and structure on the accuracy of the performance appraisal system using discrete event simulation.

From 2010 to 2016, the total number of active duty US Army personnel decreased over 17 per cent. Within the officer ranks, the Department of Defense (DoD) uses a variety of force-shaping mechanisms to downsize the services, of which the most immediate and impactful are decreased promotion rates and officer separation boards. In contrast to most civilian organizations, the Defense Officer Personnel Management Act of 1980 mandates the termination of officers twice not selected for promotion (Defense Officer Personnel Management Act, 1980). To facilitate the decrease in the number of Army personnel, the Army promotion rates to the rank of lieutenant colonel (LTC) for 2015 and 2016 were the lowest in over two decades. A key component of both promotion and officer separation boards is the analysis of officer evaluation reports (OERs), the military version of performance appraisals that contain both objective and subjective evaluation components.

Three of the four services within the DoD use a form of forced distribution when documenting officer performance or promotion potential. The US Army Evaluation Reporting System restricts raters from giving more than 49 per cent of their subordinates "most qualified" evaluations (Department of the Army, 2015). Raters within the US Navy are given a maximum number of Officer Fitness Reports (FITREPs) that can be labeled as "promote early" or "must promote" (Department of the Navy, 2016). In the US Air Force, raters submit Promotion Recommendation Forms (PRFs) on subordinate officers that have a forced distribution based on promotion zone, competitive category, and grade (Department of the Air Force, 2016). The only service that does not use a forced distribution system is the US Marine Corps, where raters score a subordinate's promotion potential, then that score is shown relative to the rater's promotion recommendations for all other subordinates (Department of the Navy, 2015). The policy constraints placed on raters within each system are just one factor that has the potential to affect the accuracy of the performance appraisals.

The biases associated with evaluating employees are well documented, particularly in management and social science literature. These biases often create a disconnect between the actual performance level of an employee and the management's perception of the employee's performance level. The absence of a forced ranking or forced distribution evaluation system can lead to rating inflation, particularly in organizations that rely heavily on performance appraisals for rewards or promotions. Bjerke *et al.* (1987) found that prior to the USA Navy implementing a forced distribution constraint on raters, over 97 per cent of Navy officers had evaluations stating they were in the top 1 per cent of their peers. However, forced distribution policy constraints, combined with the system structure, dynamics, and human behavior, induce an additional error not currently addressed in the literature. Adler *et al.* (2016) alluded to this error, noting that organizations use numerous ineffective methods for measuring employee performance, but their assessment was purely qualitative. Bartholomew and Forbes (1979) recognized the importance, and inseparable characteristics of the aggregate and the individual when modeling manpower systems, but conceded that statistical approaches are most directly relevant when analyzing the aggregate level. However, recent advances in discrete event simulation allow us to analyze the systemwide impact on individuals with unique characteristics and qualifications.

Military personnel systems have long been a subject for manpower modeling, or workforce planning, due to their size relative to most civilian organizations. Techniques for manpower modeling include dynamic programming, goal programming, Markovian models, and simulation. These techniques assist policy makers in developing strategies that match the supply of personnel with the occupational requirements. Rather than analyzing the aggregate requirements by occupation and seniority, this study seeks to determine the extent to which the current appraisal system identifies the *best* people for the available jobs. While this is often a subjective measurement, the use of discrete event simulation enables quantifying the effects of the current system and analyze future policy decisions. Our paper focuses specifically on the job turnover frequency of US Army officers, the appraisal system structure, and the effect they have on the accuracy of the performance appraisal system. For officers in the fiscal year 2016, travel expenses between duty locations due to job turnover totaled nearly $340 million (Deputy Assistant Secretary of the Army – Budget, 2017).

The frequency with which an officer moves and the accuracy of the performance appraisal system structure have an effect on an officer's career path and promotion potential. The Army Leader Development Strategy (2013) clearly states that "the Army identifies high performers and provides them with additional opportunities to broaden their perspectives." These additional opportunities include assignments, education, and training. The US Army Human Resources Command recently released the Assignment Interactive Module 2.0 (AIM 2.0) that contains a file assessment module. According to the module, officers who receive *highly qualified* or *center of mass* are viewed as performing with their peers or slightly behind their peers, whereas officers receiving *most qualified* (MQ) or *above center of mass* (ACOM), henceforth referred to as MQ/ACOM evaluations, are assessed as ahead of their peers or slightly ahead of their peers. The file strength assessment is a critical component of the officer assignment process. Furthermore, the number of MQ/ACOM evaluations an officer receives is a strong predictor for promotion. Table I shows the promotion rates for officers in the rank of major for 2015 and 2016 based on the number of MQ/ACOM evaluations received over a five-year period. Both the AIM 2.0 file strength assessment and Table I demonstrate the consequences associated with the Army Performance Evaluation System.

## Related literature

While the difference in the considered and the promoted population has been widening, the majority of recent research focused on developing models for manpower planning under uncertainty. These models are useful in determining an appropriate quantity and occupational mix for current and future authorizations. This is of particular interest to the DoD since most organizations seeking to align personnel inventory with workforce requirements treat continuation rates and workforce requirements as a given. Bastian *et al.* (2015) use stochastic goal programming techniques to solve force mix problems for Army medical specialists where continuation rates are modeled as random variables. When requirements are taken as variables, simulation-optimization has been used to determine an optimal, or near-optimal, occupational mix (Henry and Ravindran, 2005; Harper *et al.*, 2010; Zais, 2014). Other models incorporate human behavior associated with retention incentives in order to minimize the gap between personnel authorizations and inventory (Hall, 2009;

**Table I.** Promotion rates for majors based on the number of MQ or ACOM evaluations received

| | No. of MQ/ACOM Evaluations | | | | | |
|---|---|---|---|---|---|---|
| Year | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) | 5 (%) |
| 2015 | 0.0 | 2.3 | 17.3 | 73.7 | 96.9 | 98.2 |
| 2016 | 0.0 | 1.3 | 12.7 | 80.3 | 98.4 | 100.0 |

**Source:** US Army Human Resources Command

Coates *et al.*, 2010; Hall and Fu, 2015). However, Boudreau (2004) identified human resource management systems as an important area for manpower modeling. Human resource management systems play a vital role in the emerging field of talent management, aiding organizations in how to recruit, develop, and retain talented employees based on the knowledge, skills, and attributes required for current and future demands (Wardynski *et al.*, 2010). With over 89 per cent of companies linking compensation to performance appraisals, ratings have lasting effects on an employee's career, affecting organizational decisions such as staffing, promotion, and termination (Adler *et al.*, 2016). Although performance appraisal systems are an integral part of human resource management systems and talent management, very little research has been conducted to determine the effectiveness of performance appraisal systems, particularly in organizations with uncertain requirements and minimal lateral entry opportunities (Kozlowski *et al.*, 1998; Coens and Jenkins, 2000).

While not directly addressing performance appraisal systems, Dabkowski *et al.* (2011) studied the effect of multiple attrition patterns on senior leader talent. Their model assumed a bivariate normal distribution of talent with operational and non-operational aspects. Using historical promotion rates, their model quantifies the effect of different attrition rates resulting in multiple recommendations that include adjusting the timing of promotion boards and aligning officers with the operational and non-operational talent for future workforce requirements. A critical assumption in the model is that promotion boards promote officers according to their talent level, without error. In order to provide the most comprehensive picture possible, we must evaluate the system used to capture officer performance.

Evaluating officer performance is the foundation for identifying individuals for a broad range of educational and operational assignments, providing them with the background required of future leaders. Odierno (2015) asserts that "as we build cohesive teams comprised of high-performing individuals with the right talents, we build a stronger Army." GEN Odierno goes on to explain that the Army must develop a superior talent management processes that sustain the Army's competitive advantage, its leaders. Odierno (2015) states that:

> [. . .] the most important task today is to form the processes and management strategies to enable our leaders of tomorrow to thrive in the uncertain, ambiguous, and complex world they undoubtedly face.

Officer development is a multi-faceted approach that exposes officers of all ranks to a myriad of experiences and assignments that prepares them for future challenges facing military leaders. The Army Leader Development Strategy (2013) provides the framework for development in three domains: institutional, operational, and self-development. However, the realization of the goals associated with each of these domains does not, in itself, ensure that the Army is accurately documenting the talents and accomplishments of its officers to the greatest extent possible.

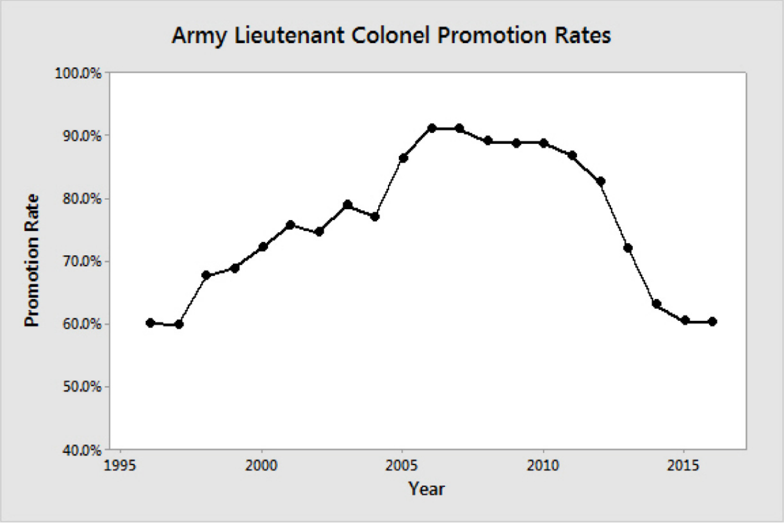
## Simulation model

Advanced analytical tools can effectively capture complex system structure and dynamics, as well as human behavior and their interactions within the US Army's performance appraisal system. In a simple way, the misidentification of high performing officers can be explained using the hypergeometric distribution given that officers are assigned to a pool from a finite population. For example, if 100 officers are separated into ten rating pools, there would be ten pools of ten officers. Each rater cannot reward their subordinates with more than 49 per cent MQ/ACOM evaluations, for a maximum of four in a pool size of ten. If the random variable $X\sim$ Hypergeometric $(K, N, n)$ where $K$ is the number of successes in a population size $N$ and $n$ represents the number of draws, we define $X\sim$ Hypergeometric (40,100,10). That is, using an ordinal ranking, there are 40 officers that fall within the rater's constraint. Assuming that raters have perfect knowledge of their subordinates' performance levels, if the 40 highest performing officers were equally distributed into the ten rating pools, all 40 would receive the appropriate MQ/ACOM rating. Conversely, the remaining 60 would appropriately receive an evaluation other than MQ/ACOM. If officers are randomly assigned into rating pools, we can determine the probability that exactly $k$ of the 40 highest performing officers are selected in the sample size of ten using the hypergeometric distribution $P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, where $k$ is the number of success drawn in the sample size $n$. For a pool size of ten, $P(X = 4) = 0.264$, i.e. the probability that a rater gives exactly four of the officers deserving an MQ/ACOM evaluation is 0.264. Furthermore, $P(X > 4) = 0.361$, meaning 36.1 per cent of the time the rater will not have enough MQ/ACOM evaluations to adequately reward his/her subordinates. If officers are sequentially assigned into pools, the parameters of the random variable $X$ are dynamic and dependent on the outcome of previous pool assignments. While the hypergeometric provides some insight into the potential misidentification of high performing officers, the performance appraisal system output is affected by many other factors, thus requiring the use of more advanced analytic techniques.

A discrete event simulation is an effective tool for traditional manpower modeling, but it also provides the means to analyze large complex systems, e.g. the US Army performance appraisal system. We first discuss the multiple simulation inputs prior to describing model validation and output analysis. Many career fields in the Army necessitate *key developmental* assignments that are prerequisites for promotion. Kane (2012) noted that raters of officers in key developmental positions are cognizant of the disproportionate impact of these evaluations and often reward these officers with MQ/ACOM evaluations as a rite of passage. To mitigate this effect, we limit the scope of our analysis to functional area officers, a subset of officers that have homogeneity of assignments due to the lack of key developmental positions. The Army promotes or accesses approximately 300 functional area officers to the rank of major each year. Our simulation model is developed based on data provided by the US Army Human Resources Command that includes the evaluations, assignment history, and promotion board outcomes of all active duty officers in the rank of major facing promotion boards in 2015 and 2016. These years reflect the most current promotion trends and rater behavior commensurate with the Army's lowest promotion rates in over 20 years, shown in Figure 1.

The flow of US Army officers through the performance appraisal system is depicted in Figure 2. Officers enter the system when they receive a promotion to the rank of major. A total of $n$ officers are then assigned into rating pools or groups of officers of the same rank with a common rater. Raters give each subordinate an evaluation in year $j$ which is a subjective measurement of the officer's performance relative to their peers within the same rating pool. The binary variable is defined as $X_{ij}$ = rating of officer $i(1,\ldots,n)$ in year $j(1,\ldots,5)$, where:

$$X_{ij}=\begin{cases}1 & \text{if officer } i \text{ receives MQ/ACOM in year } j\\ 0 & \text{if officer } i \text{ does not receive MQ/ACOM in year } j.\end{cases}$$



**Source:** US Army Human Resources Command

**Figure 1.**
Historical active duty army promotion rates to the rank of LTC

Due to high turnover and frequent moves in the military, the rated officer either remains in the same pool with probability $p$ or is reassigned into a new pool with probability $1-p$ after each evaluation. Officers exit the system when they face promotion boards after a specified time at each rank, e.g. five years. Officers exit the system with an evaluation vector $[X_{i1} \ldots X_{i5}]$ where:

$$0 \leq \sum_{j=1}^{5} X_{ij} \leq 5, \ \forall \ i = 1, \ldots, n,$$

indicating that each officer receives between zero and five MQ/ACOM evaluations over a period of five years as a major.

While the focus of this paper is on analyzing the effect of personnel movement on the performance appraisal system accuracy, it is necessary to address the procedure used in estimating rater behavior. The historical data represented in Figures 3 and 4 show that officers are more likely to receive MQ/ACOM evaluations as they increase in seniority, resulting in nearly 75 per cent of majors receiving two, three, or four MQ/ACOM evaluations over a five-year period. Dabkowski *et al.* (2011) assumed a static talent level when analyzing the effect of attrition on a senior leader talent distribution and justifies this assumption by analyzing the number of senior leaders based on their undergraduate order of merit quartile. Our model assigns each officer an initial performance percentile, $Q_i$, then sorts officers within each pool based on $Q_i'(Q_i,j,\alpha)$ without making any assumption with regard to whether the improved performance level is an actually improved performance, rater perception, or both. The unknown parameter $\alpha$ is unique to each potential sorting function $Q_i'$ and estimated using the simulation optimization routine. The goodness of fit for the sorting function $Q_i'$ is determined for each $\alpha$ parameter setting by the sum of squared error between the simulation output and the historical data. This multi-objective approach uses the following equations:

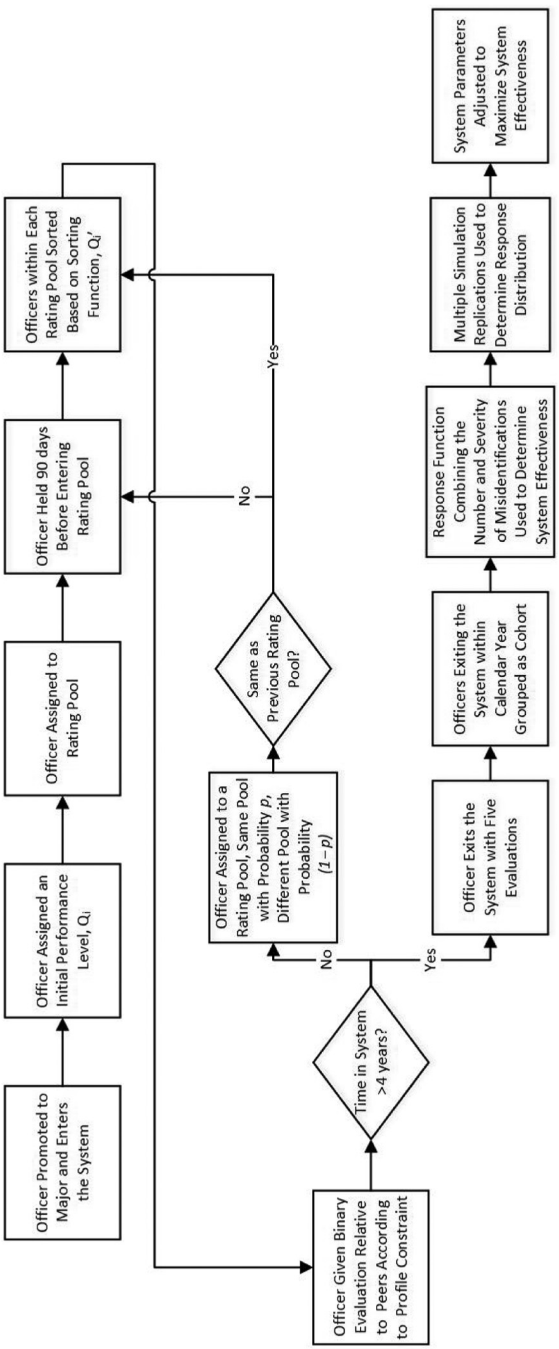
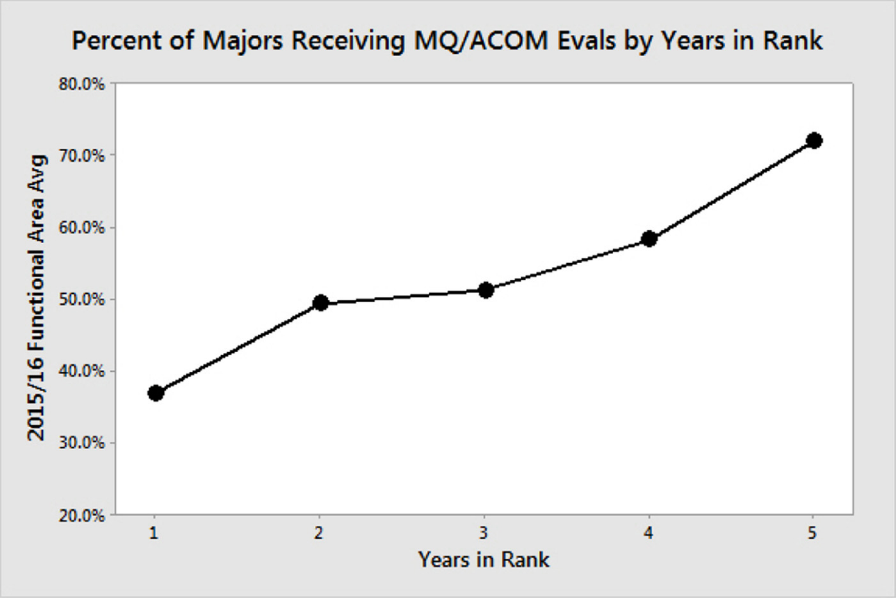
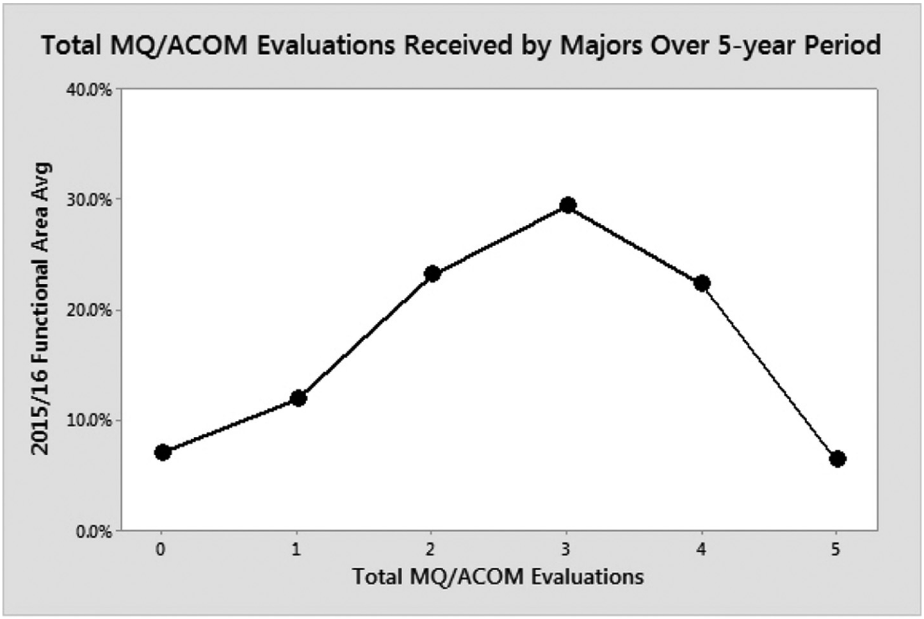

Officer Promoted to Major and Enters the System
Officer Assigned an Initial Performance Level, $Q_i$
Officer Assigned to Rating Pool
Officer Held 90 days Before Entering Rating Pool
Officers within Each Rating Pool Sorted Based on Sorting Function, $Q_i'$
Officer Given Binary Evaluation Relative to Peers According to Profile Constraint
Time in System >4 years?
No
Yes
Officer Assigned to a Rating Pool, Same Pool with Probability p, Different Pool with Probability (1−p)
Same as Previous Rating Pool?
No
Yes
Officer Exits the System with Five Evaluations
Officers Exiting the System within Calendar Year Grouped as Cohort
Response Function Combining the Number and Severity of Misidentifications Used to Determine System Effectiveness
Multiple Simulation Replications Used to Determine Response Distribution
System Parameters Adjusted to Maximize System Effectiveness

**Figure 2.**
Flow chart of the officer evaluation system simulation



**Source:** US Army Human Resources Command

**Figure 3.**
Percent of functional area majors receiving an MQ/ACOM evaluation by years in rank



**Source:** US Army Human Resources Command

**Figure 4.**
Total number of MQ/ACOM evaluations received by functional area majors over a five-year period

$$Y = \sum_{j=1}^{5} \left( \frac{\sum_{i=1}^{n} X_{ij}}{n} - B_j \right)^2 \tag{1}$$

and

$$T = \sum_{k=0}^{5} \left( \frac{\sum_{i=1}^{n} Z_{ik}}{n} - A_k \right)^2 \tag{2}$$

where:

$$Z_{ik} = \begin{cases} 1, & \text{if} \sum_{j=1}^{5} X_{ij} = k \\ 0, & \text{otherwise} \end{cases} \qquad \forall k = 0, \ldots, 5, \qquad i = 1, \ldots, n. \tag{3}$$

Equation (1) estimates the squared error between the simulation output for the percent of $n$ officers receiving an MQ/ACOM evaluation for each year as a major and the historical percentages, $B_j$, shown in Figure 4. Similarly, equation (2) calculates the squared error between the percentage of officers receiving $k$ MQ/ACOM evaluations and the historical percentages, $A_k$ shown in Figure 3, where the binary variable $Z_{ik}$ identifies the $k$ number of MQ/ACOM evaluations an officer receives in equation (3). The optimized sorting function that minimizes a linear combination of $T$ and $Y$ produces simulation output that further validates the model. A more detailed description of the optimization procedure used to estimate $Q_i{}'$ can be found in Evans *et al.* (2017).

The probability that an officer changes rating pools is estimated using OptQuest, a commercial simulation optimization engine packaged to run as an add-in within the discrete event simulation software (Simio). The annual probability $p$ that an officer changes rating pools is assessed in the optimization formulation by equations (4)-(8). We use 16.42 months as the mean amount of time ($\overline{T}$) a functional area officer spent in each assignment from 2010-2016 to search for an optimal $p$ value in the following problem:

$$\text{Minimize} \quad |T(p) - \overline{T}| \tag{4}$$

$$\text{Subject to} \quad p_j = p(1-p)^{j-1}, \qquad \forall\, j = 1, \ldots, 4 \tag{5}$$

$$p_5 = (1-p)^4 \tag{6}$$

$$T(p) = \sum_{j=1}^{5} 12j(p_j) \tag{7}$$

$$0 \leq p \leq 1 \tag{8}$$

$T(p)$ represents $E[\textit{Time in Position}]$ in equation (7), which is the expected time an officer spends in a position. $T(p)$ is a function of the annual probability that the officer changes rating pools. These probabilities are multiplied by $12j$ to calculate the expected number of months an officer spends in each assignment. Equation (5) shows the probability that an officer stays in the same assignment for one to four years. For instance, the probability that an officer stayed in the same rating pool for two years is $p(1-p)^{2-1} = (1-p)p$, which is the probability the officer did not change rating pools the first year times the probability that the officer changed rating pools the second year. The probability an officer stayed in the same rating pool for five years is shown in equation (6). The probability $p$ that minimizes equation (4) is the optimal parameter for replicating the dynamics of the Army performance appraisal system. Table II summarizes the calculation of $T(p = 0.730)$ using equations (5)-(7) and the corresponding probabilities that officers remain in the same assignment $j$ years.

For each replication, the simulation output consists of 300 officers exiting the system with an initial performance percentile, $Q_i$, and $j$ binary evaluations. A warmup period of ten years is used to fully populate the rating pools (we use the size of 15 as baseline) and ensure that officers exiting the system have all five ratings against the full complement of officers. Additionally, we run 200 replications for each scenario. This number of replications exceeds the number of replications necessary for significance during the parameter estimation phase. In Table III, the performance measures used to determine the accuracy of the performance appraisal system are the interquartile ranges of the performance percentiles of officers receiving each $k$ level of MQ/ACOM evaluations and the percentage of misidentified officers. Relatively small interquartile ranges are the result of officers with similar performance levels receiving a similar number of MQ/ACOM evaluations. The sample output in the table indicates several misidentifications. For example, $Q_4 = 0.797$ and $Q_5 = 0.845$, but $\sum_{j=1}^{5} X_{4j} = 5$ while $\sum_{j=1}^{5} X_{5j} = 4$. That is, the officer with a higher performance level received fewer MQ/ACOM evaluations.

## Results

The widths of the interquartile ranges for officers receiving each of the $k$ possible levels of MQ/ACOM evaluations and the number of misidentified officers are used for determining the accuracy of the performance appraisal system. Figure 5 shows a box plot of the simulation output for the performance percentiles of officers receiving $k$ MQ/ACOM evaluations. Overlapping interquartile ranges are indicators of errors within the system. For instance, the third quartile performance percentile of officers receiving four MQ/ACOM evaluations is greater than the first quartile of officers receiving five MQ/ACOM evaluations.

Decreasing the average rating pool size results in increased interquartile ranges. In the current US Army performance appraisal system, the average size of rating pools for majors is 15 officers. The most current revision of Army Regulation 623-3: Evaluation Reporting System requires raters to avoid the practice of *pooling*, or:

| Year ($j$) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Months ($12j$) | 12 | 24 | 36 | 48 | 60 |
| $p_j$ | 0.730 | 0.197 | 0.053 | 0.014 | 0.005 |
| $12j \times p_j$ | 8.757 | 4.733 | 1.919 | 0.619 | 0.320 |
| $T(p)$ | 16.42 | | | | |

**Note:** $p = 0.730$

**Table II.** Calculation of expected time in position for optimal $p$

> […] elevating the rating chain beyond the rater's ability to have adequate knowledge of each soldier's performance and potential, in order to provide an elevated assessment protection for a specific group (Department of the Army, 2015).

The elimination of pooling will decrease the average rating pool size. Figure 6 shows the effect of decreasing the average rating pool size to ten and five officers. Decreasing the average rating pool size results in increased interquartile ranges and performance percentile standard deviations for officers receiving each of the *k* levels of MQ/ACOM evaluations. The interquartile range of officer performance percentile when the average pool size is five is nearly double the interquartile range of officer performance percentile when the average pool size is 15 for each of the *k* MQ/ACOM levels. This implies that there is greater

**Table III.**
Sample of simulation output showing officers with varying performance levels and the evaluations received

| $i$ | $Q_i$ | $X_{i1}$ | $X_{i2}$ | $X_{i3}$ | $X_{i4}$ | $X_{i5}$ | $\sum_k X_{ij}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.272731 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0.578405 | 1 | 1 | 1 | 1 | 1 | 5 |
| 3 | 0.309698 | 0 | 0 | 1 | 1 | 1 | 3 |
| 4 | 0.797265 | 1 | 1 | 1 | 1 | 1 | 5 |
| 5 | 0.845102 | 0 | 1 | 1 | 1 | 1 | 4 |
| 6 | 0.333922 | 0 | 0 | 0 | 1 | 1 | 2 |
| 7 | 0.098697 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.988463 | 1 | 1 | 1 | 1 | 1 | 5 |
| 9 | 0.065498 | 0 | 0 | 0 | 0 | 1 | 1 |
| 10 | 0.577568 | 1 | 0 | 1 | 1 | 1 | 4 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 300 | 0.387713 | 0 | 1 | 0 | 1 | 1 | 3 |

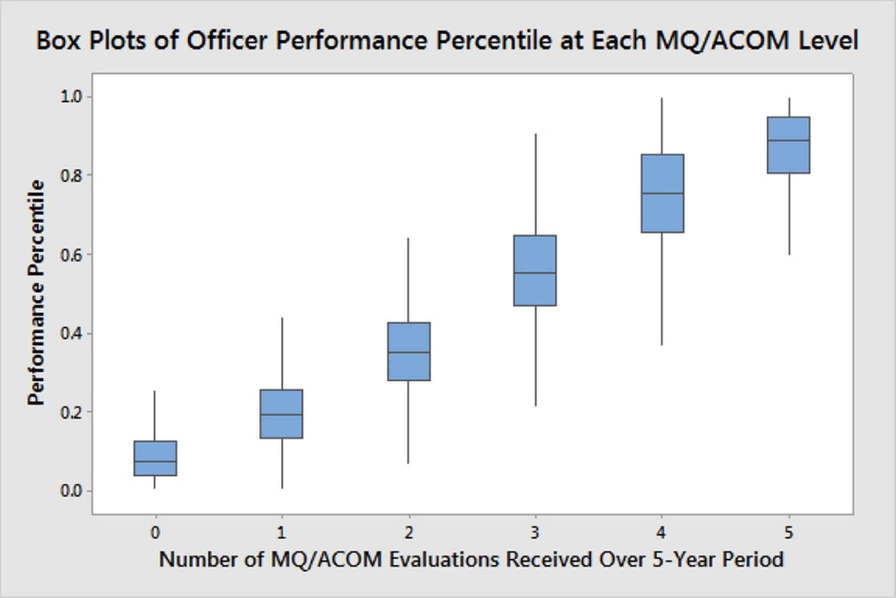


**Figure 5.**
Performance percentile distribution for officers receiving *k* MQ/ACOM evaluations

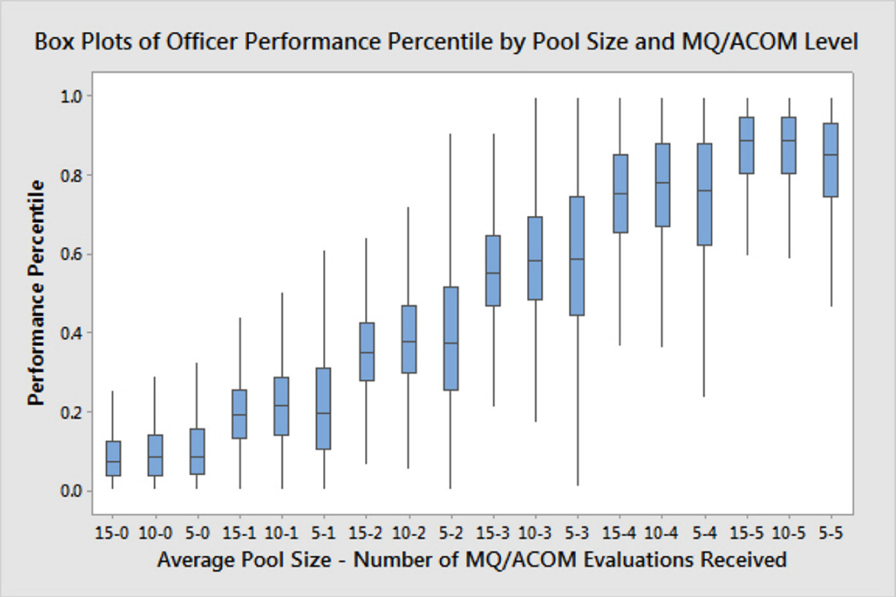


**Figure 6.** Effect of decreasing average rating Pool sizes on the performance appraisal system accuracy

variance in officer performance at each rating level for smaller pool sizes, leading to an increasing number of potential misidentifications.

Officer performance percentile interquartile range alone does not provide a comprehensive measure of performance appraisal system accuracy. The number and severity of misidentifications, combined with level $k$ at which the misidentifications occur provide additional insight into the accuracy of the system. The classification table shown in Table IV shows the officers correctly identified and misidentified at each level $k$ of MQ/ACOM evaluations with the current average rating pool of 15 officers. Table V shows the misidentifications of the current performance appraisal system and the cumulative percent of officers receiving $k$ or fewer MQ/ACOM evaluations. The values in the table refer to the percent of the total population. For each column, we classify the misidentified officers into the categories of how many MQ/ACOM evaluations the officers deserved in a perfect system. The number of evaluations an officer deserves is based off an officer's performance percentile, $Q_i$, compared to the cumulative per cent of officers receiving level $k$ of MQ/ACOM evaluations (shown in the top row of Table V). For example, in the column of officers

| | MQ/ACOM evaluations received | | | | | |
|---|---|---|---|---|---|---|
| MQ/ACOM evaluations deserved | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) | 5 (%) |
| 0 | 8.98 | 3.00 | 0.11 | 0.00 | 0.00 | 0.00 |
| 1 | 2.97 | 7.35 | 3.37 | 0.09 | 0.00 | 0.00 |
| 2 | 0.21 | 3.26 | 10.80 | 3.88 | 0.20 | 0.00 |
| 3 | 0.00 | 0.15 | 3.71 | 11.75 | 4.92 | 0.40 |
| 4 | 0.00 | 0.00 | 0.21 | 4.69 | 11.04 | 5.24 |
| 5 | 0.00 | 0.00 | 0.00 | 0.43 | 5.18 | 8.06 |

**Table IV.** Classification table of officer misidentification in the current performance appraisal system

receiving two MQ/ACOM evaluations, 3.71 per cent of the officers had a $Q_i$ greater than 0.6496 and deserved to receive three MQ/ACOM evaluations. Similarly, 0.21 per cent of the population received two MQ/ACOM evaluations but had a $Q_i$ greater than 0.8630, the equivalent of four MQ/ACOM evaluations. Conversely, this logic is also extended to the officers receiving more evaluations than they deserved. We note that the level, $k$, at which the misidentifications occur is correlated to their proportions.

The consequences associated with performance misidentification vary across each level $k$. According to Table I, officers who receive three or more MQ/ACOM evaluations were promoted at a rate greater than 70 per cent for the past two years, whereas officer receiving two or fewer MQ/ACOM evaluations were promoted at a rate of less than 20 per cent. Additionally, in terms of promotion rates, there is very little difference between officers who receive zero or one MQ/ACOM evaluations. Because of this, we have classified a subset of misidentifications as *critical misidentifications*. The critical misidentifications occur when officers deserved at least three MQ/ACOM evaluations, but received two or less, or officers who received three or more MQ/ACOM evaluations, but deserved two or less. The percent of critical misidentifications for an average pool size of 15 are marked (in italics) in Table V. Figure 7 compares the misidentifications and critical misidentifications for average pool sizes ranging from 5 to 10 officers. Increasing the average pool size results in fewer misidentifications as well as critical misidentifications.

The difference in magnitude between the misidentifications and critical misidentifications suggests that the majority of misidentifications are not egregious errors. If we consider a 3 per cent allowable error at each boundary $Q_i$ cutoff, the number of misidentifications decreases significantly. Table VI shows that for an average pool size of 15 officers, the number of misidentified officers decreases from 42.02 to 29.43 per cent when we use a 3 per cent allowable error. Additionally, the number of critical misidentifications decreases from 8.24 to 5.57 per cent. Therefore, it is a reasonable conclusion that the misidentifications in the current performance appraisal system frequently occur when an officer's performance level $Q_i$ is near the cutoff score for each $k$ level of MQ/ACOM evaluations.

In addition to analyzing the effect of average pool size on the performance appraisal system accuracy, we perform sensitivity analysis on the effect of the frequency of moves. Figure 8 shows box plots of the performance percentile distribution of the officers receiving each level ($k$) of MQ/ACOM evaluations when the time in position is one through five years. Increasing the average time in position results in a slightly wider interquartile range at each level $k$. For example, the interquartile range increases by an average of 2.2 per cent when the average time in position changes from one to two years. The interquartile range increases an average of 11.6 per cent when the average time in position increases from one to five years. Although the effect of the time in position does not appear to be significant, Figure 9 shows that there is a 20.5 per cent increase in the number of critical misidentifications between an average time in position of one year (8.18 per cent) and an average time in position of five

**Table V.**
Percent of officers misidentified in the current performance appraisal system

| | MQ/ACOM evaluations received | | | | | |
|---|---|---|---|---|---|---|
| Percent of officers | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) | 5 (%) |
| Cumulative % | 12.16 | 25.92 | 44.12 | 64.96 | 86.30 | |
| Deserved 1 More | 2.97 | 3.26 | *3.71* | 4.69 | 5.18 | |
| Deserved 2+ More | 0.21 | *0.15* | *0.21* | 0.43 | | |
| Deserved 1 Less | | 3.00 | 3.37 | *3.88* | 4.92 | 5.24 |
| Deserved 2+ Less | | | 0.11 | *0.09* | *0.20* | 0.40 |
| Misidentifications | 42.02 | | | | | |
| Critical Misidentifications | *8.24* | | | | | |

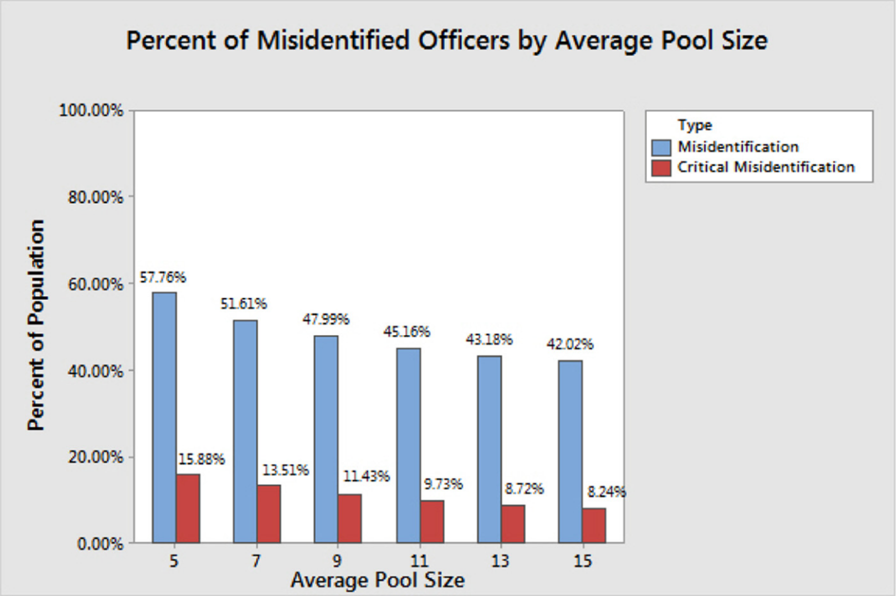


**Figure 7.** Percent of officer misidentifications and critical misidentifications when varying the average pool sizes

**Table VI.** Percent of officers misidentified in the current performance appraisal system with 3% allowable error

| Average Pool Size | Pool Size 15 (3% Allowable Error) | | | | | |
|---|---|---|---|---|---|---|
| MQ/ACOM Evaluations | 0 (%) | 1 (%) | 2 (%) | 3 (%) | 4 (%) | 5 (%) |
| Cumulative % | 12.16 | 25.92 | 44.12 | 64.96 | 86.30 | |
| Deserved 1 More | 1.79 | 2.10 | *2.70* | 3.72 | 3.62 | |
| Deserved 2+ More | 0.10 | *0.05* | *0.11* | 0.30 | | |
| Deserved 1 Less | | 1.69 | 2.10 | *2.52* | 3.75 | 4.42 |
| Deserved 2+ Less | | | 0.05 | *0.03* | *0.16* | 0.22 |
| Misidentifications | *29.43* | | | | | |
| Critical Misidentifications | *5.57* | | | | | |

years (9.86 per cent). This indicates more misidentification occurrences when officers stay in positions for longer periods of time.

## Discussion

Decreasing the average pool size results in a non-linear increase in the number of misidentified officers. The most recent revision of *Army Regulation 623-3: Evaluation Reporting System* mandates that ratings must be delegated to the lowest level possible, the resulting smaller rating pools create a greater number of misidentifications (Department of the Army, 2015). There needs to be a balance between rating pools that are small enough for raters to have an intimate knowledge of their subordinates' performance, but large enough to provide the flexibility to adequately reward high-performing individuals. Figure 7 illustrates that decreasing the average rating pool size from the current size of 15 to 10 results in a 19.5 per cent increase in the number of critical misidentifications. However, when the average rating pool size is decreased from 15 to 5, the number of critical

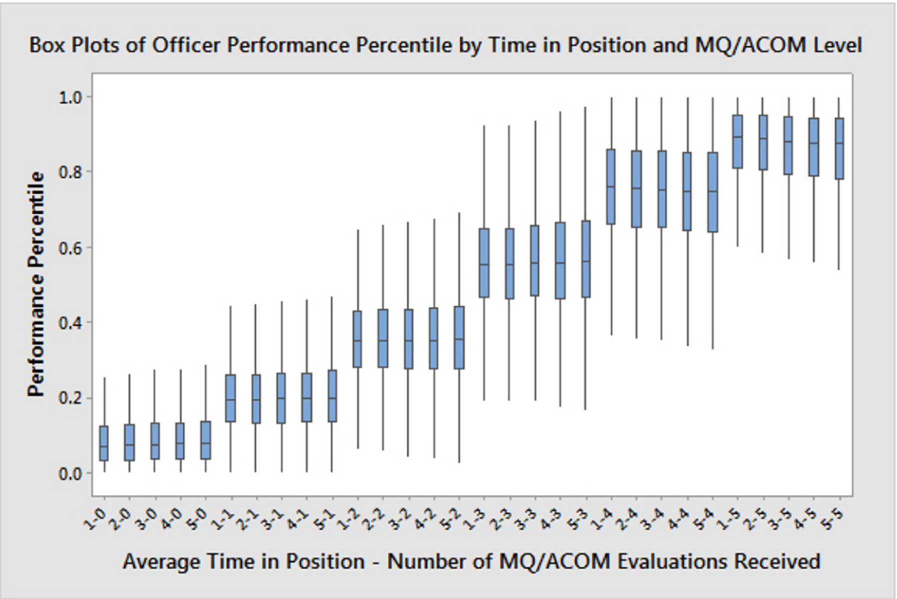


**Figure 8.**
Effect of changing the average time in position on the performance appraisal system accuracy

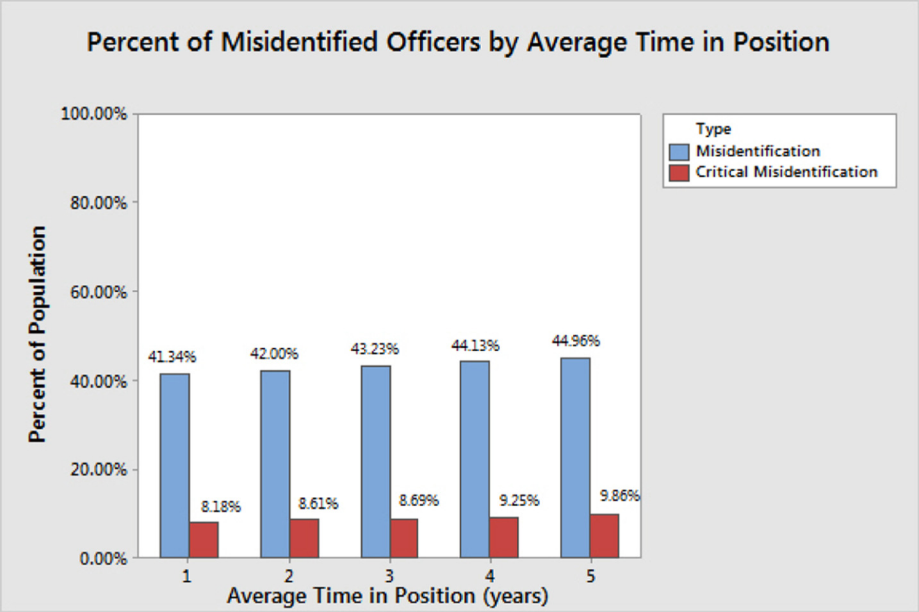


**Figure 9.**
Percent of officer misidentifications and critical misidentifications when varying the average time in a position

misidentifications nearly doubles. Hence, moderate decreases in sizes of rating pools appear to be acceptable, whereas drastic decreases would have the unintended consequence of a large decrease in performance appraisal system accuracy.

Decreasing the frequency of moving an officer can serve as an effective cost-cutting measure, but serves in turn as a hindrance to professional development. In 2016, the Army spent an average of nearly $19,000 per move for operational and rotational travel. Individual officers spend an average of 16.42 months in each position, and 83.6 per cent of every change in position results in a change in duty location, accounting for $38,550 in moving expenses to the average major between assignments over a five-year period (Deputy Assistant Secretary of the Army – Budget, 2017). Keeping officers in an assignment for a longer limits the breadth of experiences critical to developing leaders of the future (Odierno, 2015). However, in terms of performance appraisal accuracy, moderate increases in the average time in position do not greatly affect the number of performance misidentifications.

This study provides a framework for assessing multiple factors and the extent to which they limit or enhance the accuracy of an organization's forced distribution performance appraisal system. In addition to the aforementioned system structure and dynamics, our simulation approach is used to effectively analyze the effects of human behavior and evaluation policy. Future research will include utilizing simulation optimization routines to analyze multiple combinations of performance appraisal system inputs and regulations in order to optimize system effectiveness. This proposed research would allow an organization's leadership to devise an evaluation policy that minimizes the unintended possible consequences associated with the proposed changes.

## References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R. and Pulakos, E.D. (2016), "Getting rid of performance ratings: genius or folly? A debate", *Industrial and Organizational Psychology*, Vol. 9 No. 2, pp. 219-251.

Army Leader Development Strategy (2013), available at http://usacac.army.mil/core-functions/leader-development (accessed 17 August 2017).

Bartholomew, D.J. and Forbes, A.F. (1979), *Statistical Techniques for Manpower Planning*, John Wiley and Sons, New York, NY.

Bastian, N.D., McMurry, P., Fulton, L.V., Griffin, P.M., Cui, S., Hanson, T. and Srinivas, S. (2015), "The AMEDD uses goal programming to optimize workforce planning decisions", *Interfaces*, Vol. 45 No. 4, pp. 305-324.

Bjerke, D.G., Cleveland, J.E., Morison, R.F. and Wilson, W.C. (1987), "Officer fitness report evaluation study", Technical Report (NPRDC TR 88-4), Navy Personnel Research and Development Center, San Diego, CA.

Boudreau, J.W. (2004), "Organizational behavior, strategy, performance, and design in management sciences", *Management Science*, Vol. 50 No. 11, pp. 1463-1476.

Coates, H.R., Silvernail, T.S., Fulton, L.V. and Ivanitskaya, L. (2010), "The effectiveness of the recent army captain retention program", *Armed Forces & Society*, Vol. 3 No. 1, pp. 5-18.

Coens, T. and Jenkins, M. (2000), *Abolishing Performance Appraisals: Why They Fail and What to Do Instead*, Berrett-Koehler Publishers, San Francisco, CA.

Dabkowski, M.F., Huddleston, S.H., Kucik, P. and Lyle, D.S. (2011), "Shaping senior officer talent: using a multi-dimensional model of talent to analyze the effect of personnel management decisions and attrition on the flow of army officer talent throughout the officer career model", in Jain, S., Creasy, R.R., Himmelspach, J., White, K.P. and Fu, M. (Eds), *Proceedings of the 2011 Winter Simulation Conference*, *Institute of Electrical and Electronic Engineers*, *Piscataway, NJ*, pp. 2466-2477.

Defense Officer Personnel Management Act (1980), *Public Law 96-513, 96th Congress*, available at: www.gpo.gov/fdsys/pkg/STATUTE-94/pdf/STATUTE-94-Pg2835.pdf (accessed 22 June 2017).

Department of the Air Force (2016), *Air Force Instruction 36-2406: Officer and Enlisted Evaluation Systems*, available at: http://static.e-publishing.af.mil/production/1/af_a1/publication/afi36-2406/afi36-2406.pdf (accessed 25 June 2017).

Department of the Army (2015), *Army Regulation 623-3: Evaluation Reporting System*, available at: www.apd.army.mil/epubs/DR_pubs/DR_a/pdf/web/r623_3.pdf (accessed 2 April 2017).

Department of the Navy (2015), *Marine Corps Order 1610.7: Performance Evaluation System*, available at: www.marines.mil/Portals/59/Publications/MCO%201610.7.pdf (accessed 28 June 2017).

Department of the Navy (2016), *Bureau of Personnel Instruction 1610.10D: Navy Performance Evaluation System*, available at: www.public.navy.mil/bupersnpc/reference/instructions/BUPERSInstructions/Documents/1610.10D.pdf (accessed 26 June 2017).

Deputy Assistant Secretary of the Army – Budget (2017), "Fiscal year (FY) 2018 president's budget submission", *Army Military Personnel Justification Book*, pp. 124-129, available at: www.asafm.army.mil/documents/BudgetMaterial/fy2018/mpa.pdf (accessed 30 August 2017).

Evans, L.A., Bae, K.-H.G. and Roy, A. (2017), "Single and multi-objective parameter estimation of a military personnel system via simulation optimization", in Chan, W.K.V, D'Ambrogio, A., Zacharewicz, G., Mustafe, N., Wainer, G. and Page, E. (Eds), *Proceedings of the 2017 Winter Simulation Conference*, *Institute of Electrical and Electronic Engineers*, *Piscataway, NJ*, pp. 4058-4069.

Hall, A.O. (2009), "Simulating and optimizing: military manpower modeling and mountain range options", PhD, University of Maryland, College Park.

Hall, A.O. and Fu, M.C. (2015), "Optimal army officer force profiles", *Optimization Letters*, Vol. 9 No. 8, pp. 1769-1785.

Harper, P.R., Powell, N.H. and Williams, J.E. (2010), "Modelling the size and skill-mix of hospital nursing teams", *Journal of the Operational Research Society*, Vol. 61, pp. 768-779.

Henry, T.M. and Ravindran, A.R. (2005), "A goal programming application for army officer accession planning", *INFOR: Information Systems and Operational Research*, Vol. 43 No. 2, pp. 111-119.

Kane, T. (2012), *Bleeding Talent*, Palgrave Macmillan, New York, NY.

Kozlowski, S.W.J., Chao, G.T. and Morrison, R.F. (1998), "Games raters play: politics, strategies, and impression management in performance appraisal", in Smither, J.W. (Ed.), *Performance Appraisal: State of the Art in Practice*, Jossey-Bass Publishers, San Francisco, CA, pp. 163-205.

Odierno, R.T. (2015), "Leader development and talent management: the army competitive advantage", *Military Review*, July-August 2015, pp. 9-15.

Wardynski, C., Lyle, D.S. and Colarusso, M.J. (2010), "Towards a US army officer corps strategy for success: retaining talent", *Strategic Studies Institute: Officer Monograph Series*, Vol. 2.

Zais, M.M. (2014), "Simulation-optimization, Markov chain and graph coloring approaches to military manpower modeling and deployment sourcing", PhD, University of Colorado.

**Corresponding author**
Lee Evans can be contacted at: lee.evans@louisville.edu