5-2018

Simulation-based analysis and optimization of the United States Army performance appraisal system.

Lee A. Evans
University of Louisville

SIMULATION-BASED ANALYSIS AND OPTIMIZATION OF THE UNITED STATES ARMY PERFORMANCE APPRAISAL SYSTEM

Lee A. Evans
B.S., United States Military Academy, 2000
M.S., Georgia Institute of Technology, 2009

A Dissertation Submitted to the Faculty of
the J.B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Industrial Engineering

Department of Industrial Engineering
University of Louisville
Louisville, Kentucky

May 2018

SIMULATION-BASED ANALYSIS AND OPTIMIZATION OF THE UNITED STATES ARMY PERFORMANCE APPRAISAL SYSTEM

Lee A. Evans
B.S., United States Military Academy, 2000
M.S., Georgia Institute of Technology, 2009

Dissertation Approved on

April 13, 2018

By the following Dissertation Committee

Dr. Ki-Hwan G. Bae, Chair

Dr. Lihui Bai

Dr. Erin Gerber

Dr. Lee Bewley

ACKNOWLEDGMENTS

My sincere gratitude goes to my advisor, Dr. Ki-Hwan Bae, for his mentorship and guidance throughout this process. I would also like to thank my dissertation committee of Dr. Lihui Bai, Dr. Erin Gerber, and Dr. Lee Bewley for generously sharing their time and ideas. This research would not have been possible without the support from the United States Army Human Resources Command. I would like to thank Mr. David Martino for his willingness to provide all of the data used to analyze the Army's performance appraisal system. His intellectual curiosity has forced the Officer Professional Management Directorate to take a critical view of itself, resulting in an organizational culture that demands continuous improvement. Mr. Martino never lets a subordinate forget that behind every number is a person, a story, and a family; a mantra that has stuck with me throughout this study. Additionally, I would like to thank Mr. Ike Zeitler, Ms. Teresa Monroe, and MAJ Nick Paul of the Officer Readiness Division for the countless hours spent querying databases in support of this dissertation.

I have been extremely fortunate to learn from wonderful public educators; the most influential being my parents, Bill and Linda Evans, who combine for over 50 years experience teaching at the high school level. From flight school to graduate school, their support and encouragement has pushed me to expand my horizons and has made me a better person. Finally, I would like to thank my wife, Kari,

and my children, Elin, Brody, and Grant, for their love and continuous support throughout my time at the University of Louisville and during our entire 18-year journey in the Army.

ABSTRACT

SIMULATION-BASED ANALYSIS AND OPTIMIZATION OF THE UNITED STATES ARMY PERFORMANCE APPRAISAL SYSTEM

Lee A. Evans

April 13, 2018

From 2010 to 2016, the total number of active duty United States Army personnel decreased by over 17%. The Department of Defense uses a variety of instruments to downsize the services, of which the most immediate and impactful is through decreased promotion rates. The Defense Officer Personnel Management Act of 1980 mandates the termination of officers twice not selected for promotion. As such, the promotion rates to the rank of lieutenant colonel (LTC) for 2015 and 2016 were the lowest over the past two decades. Central to each promotion board is the analysis of officer evaluation reports (OERs), the military version of performance appraisals.

The biases associated with evaluating employees are well documented, particularly in management literature. These biases can often create a disconnect between the actual performance level of an employee and the management's perception of the employee's performance level. The performance appraisal system in the United States Army is a forced distribution system that restricts the number of above average evaluations raters are allowed to give subordinates. This structure, combined with human behavior and system dynamics, creates an additional bias not currently addressed in literature.

Military personnel systems have long been the subjects 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 with matching the supply of personnel with the available jobs. Rather than analyzing the aggregate requirements by occupation and seniority, this study determines the extent to which the current system promotes the *best* people into the available jobs. While this is often a subjective measurement, the use of discrete event simulations allows us to quantify the effects of the current system and analyze future policy decisions.

In this dissertation, a discrete event simulation framework is considered to replicate the dynamics, structure, and regulatory constraints placed on the officers in the U.S. Army. Using performance appraisal data provided by the United States Army Human Resources Command, we create a multi-objective response function in order to quantify the human behavior associated with evaluating subordinates. We are able to minimize the squared error of our system output with the multi-objective response function using simulation-optimization techniques. Utilizing simulation-optimization techniques for model validation enables estimating unknown input parameters, such as human behavior, based on historical data. Furthermore, the model allows users to analyze the effects of current constraints on the evaluation system and the effects of proposed personnel policy changes.

The effectiveness of the performance appraisal system is based on its ability to

accurately evaluate the officers' performance levels. The model output is analyzed by the number of misidentified individuals and the severity of the misidentification. An initial analysis showed that 20.07% of the officers in the system do not receive as many above average evaluations as their performance level warrants. Additionally, structural changes such as decreasing the average number of a rater's subordinates from fifteen to five increases the number of misidentified personnel by 59.86%. Ranking and selection methods that include the Kim Nelson (KN) and the Nelson, Swann, Goldsman, Song (NSGS) procedures assists in determining the optimal combination of input parameters such as forced distribution constraints placed on raters, frequency of moves, number of subordinates assigned to each rater, and rater behavior.

The simulation will serve as a tool for policy analysis to recommend policies and behavior that maximizes the extent to which the performance appraisal system accurately identifies the most qualified employees. Consequently, the results demonstrate broad applicability of simulation-optimization in the field of manpower modeling and human resource management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Despite personnel costs comprising nearly half of the Department of Defense's $585 billion annual budget, active duty United States Army personnel levels have rapidly decreased as a result of budgetary constraints and congressional authorizations (Office of the Under Secretary of Defense (2015)). According to the Defense Manpower Data Center, the number of active duty U.S. Army personnel has decreased by nearly 17% since 2010 (see Figure 1.1). The Department of Defense (DoD) employs a number of techniques such as decreased accessions, involuntary retirement, separation boards, fewer re-enlistment opportunities, and lower promotion rates to facilitate this decrease in personnel strength levels. Figure 1.2 shows the corresponding decrease in active duty promotion rates to the rank of lieutenant colonel for the Army Competitive Category (ACC). The ACC includes all branches of the Army with the exception of medical officers, Judge

Advocate General (JAG) officers, and chaplains. Given the specificity of these occupations, non-ACC officers go through a separate accession and promotion process. Henceforth, any reference to evaluation, promotion, attrition, or assignment applies strictly to ACC officers. Traditionally, military manpower modeling has focused on retaining an optimal mix of skills and experience in order to meet a future demand with specified degrees of uncertainty. As an organization with minimal lateral entry opportunities, this approach allows the DoD to set policy and determine retention incentives for future requirements. While traditional manpower modeling aids decision makers in determining the number and occupational distribution of the organization's employees, the emerging field of talent management seeks to acquire, promote, and develop the right candidates for each job requirement.

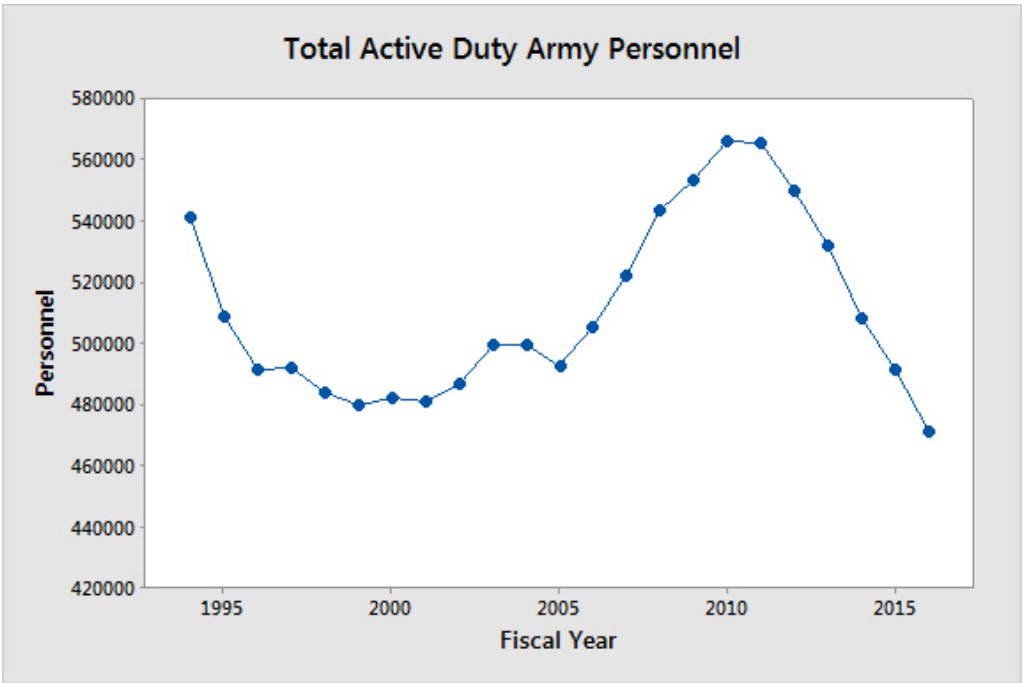


Figure 1.1: U.S. Army active duty personnel strength from 1994 to 2016 (Source: Defense Manpower Data Center)

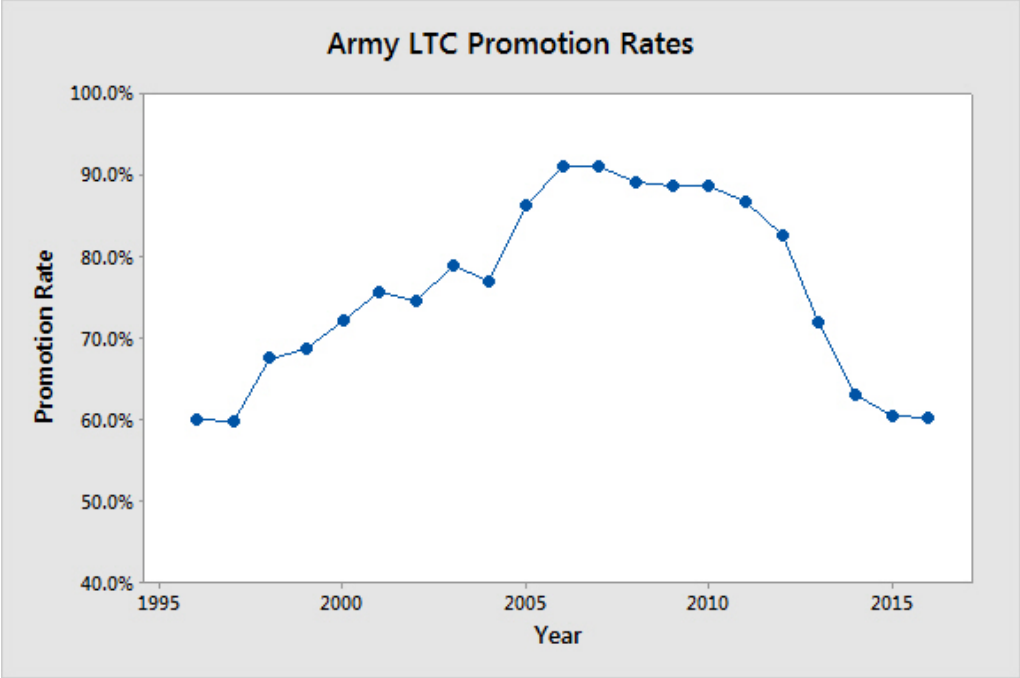


Figure 1.2: U.S. Army active duty lieutenant colonel promotion rate from 1996 to 2016 (Source: U.S. Army Human Resources Command)

Talent management has become a topic of interest for numerous organizations within the DoD. The U.S. Army Deputy Chief of Staff for Personnel (G1), the U.S. Army Human Resources Command, the U.S. Army Cadet Command, and the Office of Economic and Manpower Analysis all have programs or teams dedicated to studying and implementing talent management within their organizations. The Office of Economic and Manpower Analysis defined talent management as the process that aligns systematic planning with implementation to meet the current and future talent demands (Wardinski et al. (2010)). This process integrates acquisition, development, employment, and retention of Army personnel. A considerable amount of attention has been given to talent management in the civilian sector as well. McDonnell et al. (2017) noted that despite the volume of research devoted to talent management, the nature of the field remains disjointed and there is an

increasing need for nuanced methodological approaches. One nuanced approach is the use of discrete-event simulation to gain insight into the behavior of the performance appraisal system.

U.S. Army performance appraisals have undergone nine major revisions since the introduction of the *forced choice* officer evaluation system in July of 1947. Upon inception, the forced choice system established rating pools of 12 to 40 officers where raters would sequentially evaluate the highest and lowest performing officers within their assigned pool. Once an officer was rated as the highest or lowest of the pool, he was removed from the next iteration of the rater's selections. The rater would repeat the process until all of the officers under consideration had been rated, creating a $\{1,\ldots,n\}$ ranking of the officers (Sisson (1948)). This ranking was used for the purposes of promotion and reassignment of officers. Officers not selected for promotion were separated from the service, a practice eventually codified by the Defense Officer Personnel Management Act (DOPMA), passed by Congress on December 12, 1980. The Defense Officer Personnel Management Act provides guidelines for the number of officers at each rank as a function of the total number Army personnel (Rostker et al. (1993)). Presently, officers are evaluated by promotion boards in cohort year groups, generally determined by their number of years of service as an officer. Any officer not selected for promotion to the next rank is forced to leave the service. There are rare exceptions to the separation mandate, but DOPMA states that any exception is to be used sparingly, generally for officers with difficult-to-replace, unique skill sets. The guidelines set forth by DOPMA create the diminishing rank structure shown in Figure 1.3, commonly

referred to as a pyramid structure due a decreasing required number of officers at higher ranks.

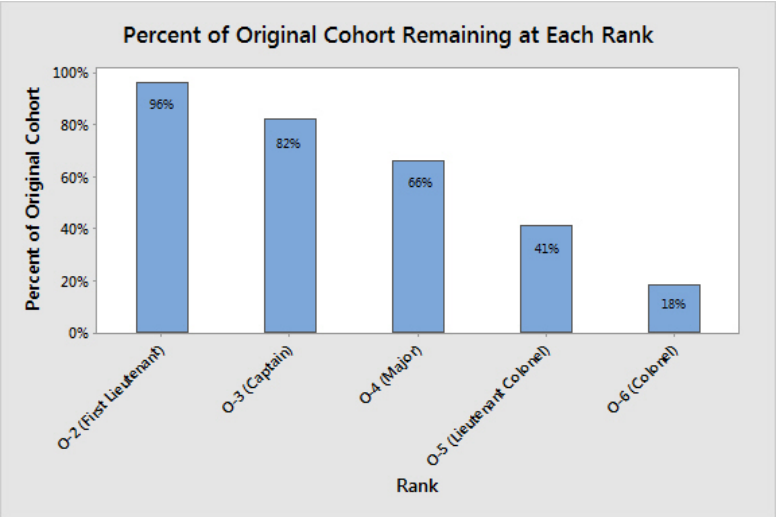


Figure 1.3: Promotion induced attrition pattern prescribed in DOPMA (from Rostker et al., 1993).

Officers in the U.S. Army face numerous promotion boards throughout their careers. Unlike many organizations, if an officer is not selected for promotion, he/she is forced out of the service no later than the first day of the seventh month after the President of the United States has approved the board results (Rostker et al. (2011)). The impact of the DOPMA separation mandate is severe due to the military's cliff vesting retirement system. That is, officers who leave service with less than 20 years do so with no retirement benefits. Given the gravity of promotion board results, officer performance appraisals, or officer evaluation reports (OERs), are critical in identifying and promoting high-performing individuals. Promotion boards consist of 17 general officers with little or no personal knowledge of the officers presented at the board. Each general officer scores the file of each officer presented at the board on a scale of zero to six, in half point increments. The

scores are aggregated, officers are placed on a $\{1,\ldots,n\}$ list, and a percentage of the officers are selected for promotion based on authorizations. The officer file consists of a one-page summary known as the officer record brief (ORB), awards, transcripts, any adverse action or letters of reprimand, and officer evaluations.

The flow of U.S. Army officers through the performance appraisal system is depicted in Figure 1.4. Officers enter the system when they receive a promotion. The officers are then assigned into *rating pools*, or groups of officers of the same rank with a common rater. Annually, raters give each subordinate officer an evaluation that is a subjective measurement of the officer's performance relative to their peers within the same rating pool. Due to high turnover and frequent moves in the military, following the evaluation the rated officer either remains in the same pool, or is reassigned into a new pool. Officers face promotion boards after a specified time at each rank, five years in the case of Figure 1.4. After each member of the promotion board scores the officers' files, the aggregated scores are used to generate an order of merit list. The officer's standing on the order of merit list ultimately determines who is promoted to the subsequent rank and who is forced to leave military service.

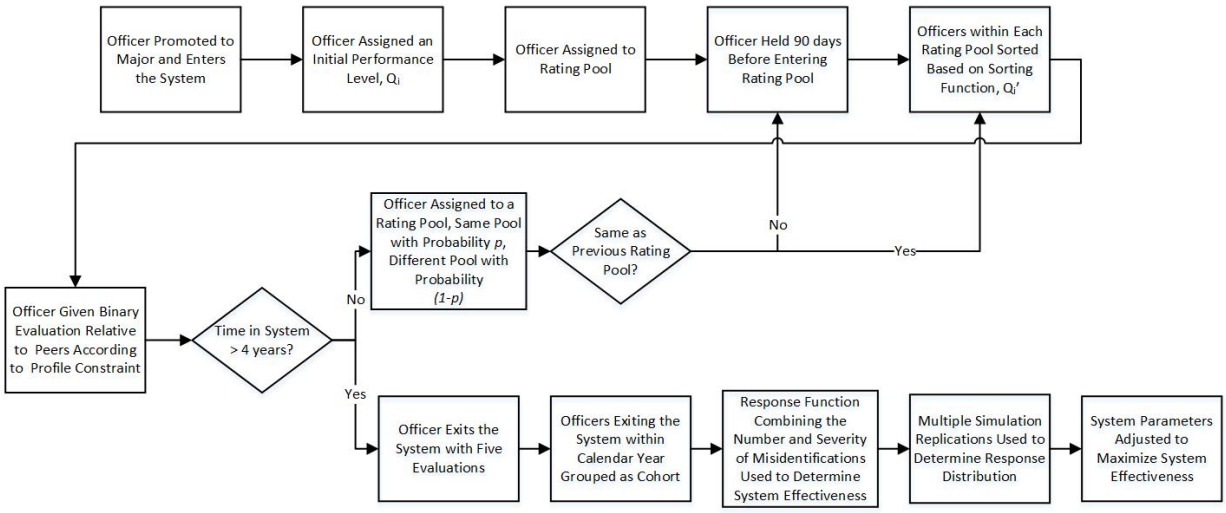
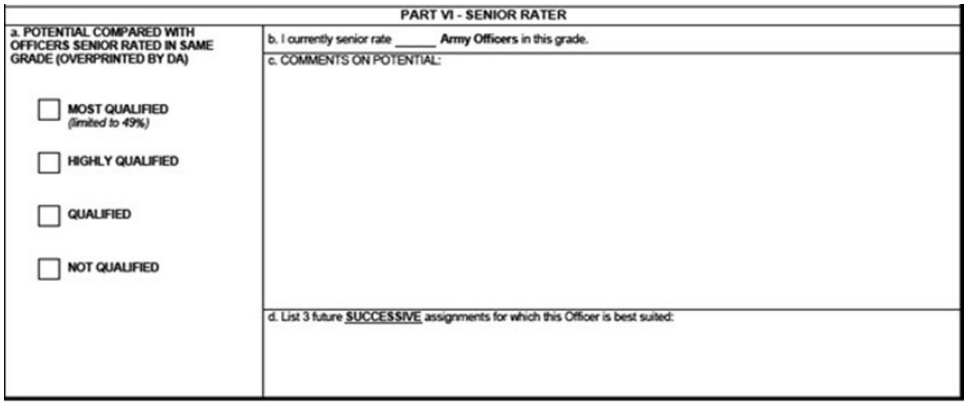


Figure 1.4: Flow chart of the simulated U.S. Army officer performance appraisal system

Due to the regulatory constraints placed on the system, human behavior, and the dynamics of frequent moves, the performance appraisal system is prone to errors. By simulating system dynamics and rater behavior, we are able to estimate the misidentification of performance in the current performance appraisal system. Simulation-optimization allows us to evaluate changes in controllable parameters in order to reduce this number of misidentifications.

Forced distribution performance appraisal systems, widely abandoned in the civilian sector, constrain raters and induce error in identifying the top performing individuals within an organization. By regulation, raters in the Army are allowed to give up to 49% top evaluations, commonly referred to as "top block", "above center of mass (ACOM)", or "most qualified" evaluations (Department of the Army (2015)). An excerpt of the Department of the Army Form 67-10-2, the Army Field Grade Officer Evaluation Report, is shown in Figure 1.5. This shows that the rater is required to check the block that corresponds to the rated officer's

potential, as compared to the potential of the other officers of the same rank. The top block, or "most qualified", is limited to no more than 49% of the total evaluations given by the rater.


PART VI - SENIOR RATER

a. POTENTIAL COMPARED WITH OFFICERS SENIOR RATED IN SAME GRADE (OVERPRINTED BY DA)

☐ MOST QUALIFIED (limited to 49%)

☐ HIGHLY QUALIFIED

☐ QUALIFIED

☐ NOT QUALIFIED

b. I currently senior rate _____ Army Officers in this grade.

c. COMMENTS ON POTENTIAL:

d. List 3 future SUCCESSIVE assignments for which this Officer is best suited:

Figure 1.5: Excerpt from Department of the Army Form 67-10-2, Field Grade Officer Evaluation Report (Source: Department of the Army Regulation 623-3: Evaluation Reporting System)

During the past two years with the rapid decrease in promotion rates to lieutenant colonel, data shows that the number of ACOM evaluations is a strong indicator of whether or not an officer is selected for promotion (see Table 1.1). The quasi-complete separation of the data prohibits the use of logistic regression, but the statistics show that officers with three or more ACOM evaluations have a very high promotion rate, while officers with two or fewer ACOM evaluations are rarely selected for promotion. Within the DoD, the consequences of misidentifying the best candidates for promotion carries dire consequences. From an organizational perspective, a suboptimal outcome of a promotion board compromises national security capabilities and the ability to confront both state and non-state actors. On an individual level, officers who have faithfully served their country and deserve to continue service are forced out of the military, often times with no retirement

benefits.

Table 1.1: LTC promotion rates by number of ACOM evaluations

		Number of ACOM Evaluations					
		0	1	2	3	4	5
Promotion Rate	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%

1.2 Research Motivation

"Leadership must move from the performance appraisal system to the appraisal of the performance of the system"

— Ronald D. Moen, *Quality Progress*

In practice, U.S. Army officers have three opportunities for promotion to each successive rank: below the zone, primary zone, and above the zone. A below the zone promotion occurs one year earlier than the majority of the officers' peers, those commissioned as officers in the same time frame. Given the limited number of officers selected for promotion below the zone, only officers not selected for promotion in the primary zone and above the zone boards face separation from military service. The majority of promotions occur during the primary zone. Officers not selected for promotion during the primary zone are afforded one additional opportunity for promotion, the above the zone board. While each officer has up to three opportunities for promotion, Figure 1.6 shows that historically, the percentage of officers promoted in any board other than the primary zone

is extremely low. Therefore, primary zone promotions provide the largest sample when analyzing the effectiveness of the U.S. Army's performance appraisal system.

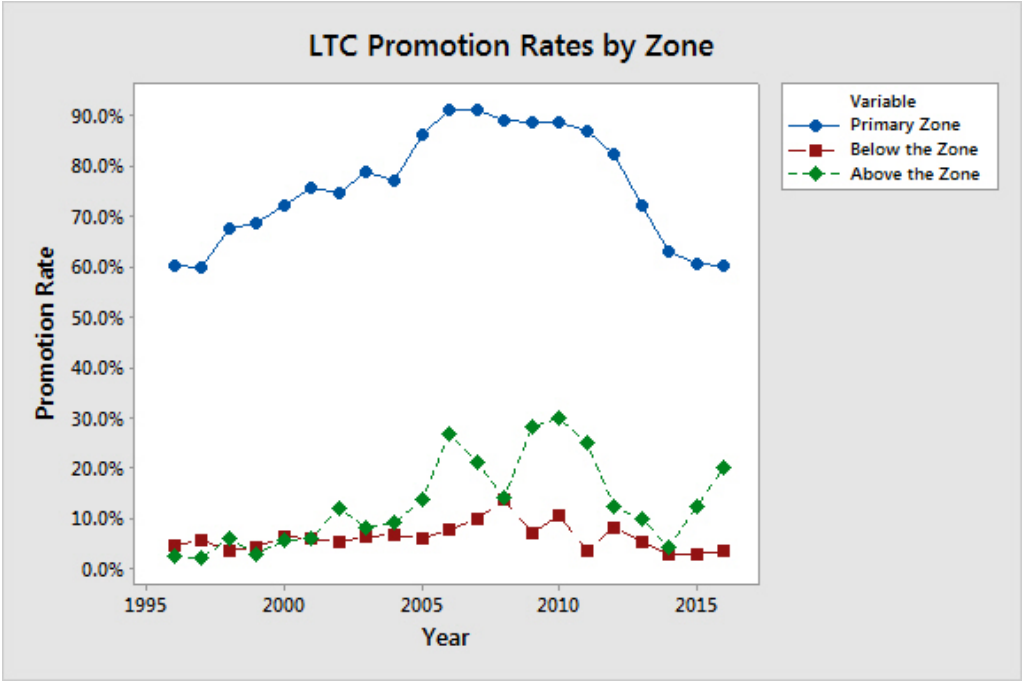


Figure 1.6: Promotion rates to the rank of lieutenant colonel by zone of consideration (Source: U.S. Army Human Resources Command)

Within the primary zone, there has been an increasing gap between those officers considered for promotion and the officers subsequently selected for promotion (see Figure 1.7). The difference between the solid and dashed lines represents the number of officers considered, but not selected for promotion. In 2015, there were 588 officers in the primary zone not selected for promotion to lieutenant colonel, while in 2016 there were 609. What remains unknown is the extent to which the U.S. Army selected the 1,816 most qualified officers for promotion and not selected the 1,197 officers least qualified for promotion over the two-year period from 2015 to 2016.

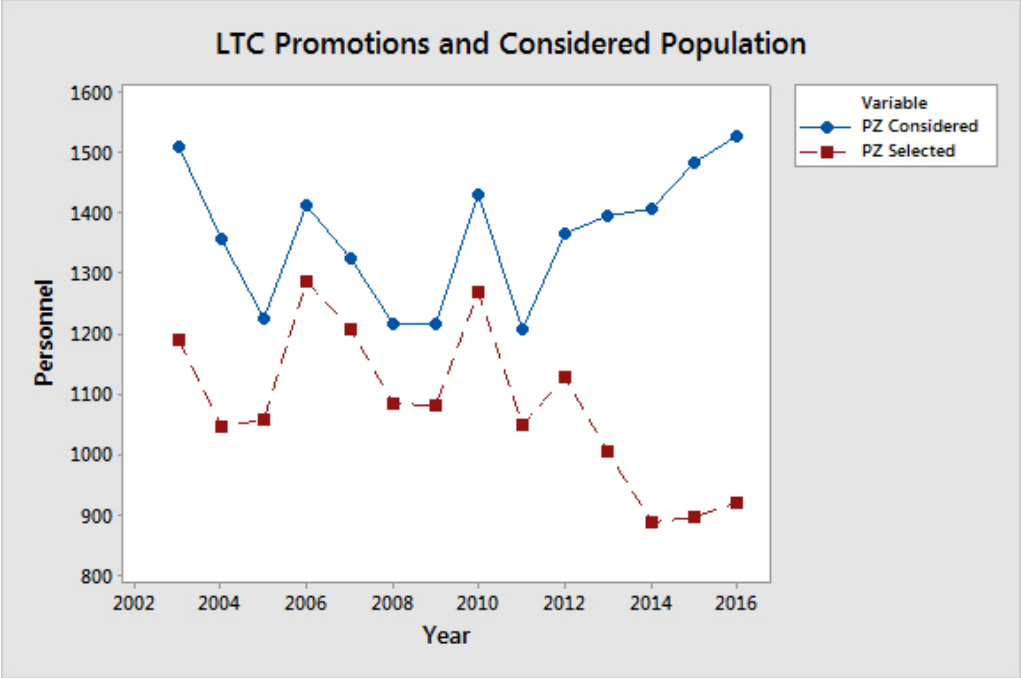


Figure 1.7: Considered and selected populations to the rank of LTC (Source: U.S. Army Human Resources Command)

While the gap between considered and promoted population has begun to widen, the majority of current research has been on developing models for manpower planning under uncertainty. These models are useful in determining an appropriate quantity and occupational mix for future authorizations. This is of particular interest to the DoD because most organizations seeking to align personnel inventory with workforce requirements treat workforce requirements as a given. When requirements are taken as variables, simulation-optimization has been used to determine an optimal, or near-optimal, occupational mix (Henry et al. (2005), Harper et al. (2010), and Zais (2014)). However, other models incorporate human behavior associated with retention incentives in order to minimize the gap between personnel authorizations and inventory (Hall (2009), Coates et al. (2010)). Boudreau (2004) identified human resource management systems as

an important area for manpower modeling. Of particular interest is the emerging field of talent management that aids 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)). While 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)). Accuracy of the performance appraisal system is critical to leader development. Odierno (2015) stated that "as we build cohesive teams comprised of high-performing individuals with the right talents, we build a stronger Army." The motivation of this research is to bridge the gap between talent management and manpower modeling through the use of discrete-event simulation by determining how the performance appraisal system and human behavior compromise the ability to accurately evaluate personnel within an organization.

This research contributes to the field of manpower modeling by gaining insight into complex human resource management systems, illustrated through the following techniques: (1) discrete-event simulation model development that incorporates the existing structure and behavior of the U.S. Army performance appraisal system; (2) model validation using historical data provided by the U.S. Army Human Resources Command; and (3) using simulation-optimization techniques to adjust controllable parameters in order to improve upon the existing performance appraisal system. Through this research, we develop techniques to

determine the misidentification of high performing personnel that is a result of regulatory constraints, system dynamics, and human behavior. Finally, we apply simulation-optimization techniques, leading to policy recommendations on how to improve the system and analyze future personnel policy decisions.

1.3 Organization of this Dissertation

This dissertation proposal is organized as follows. In Chapter 2, we review the literature on models for military modeling and simulation, performance appraisal systems, military policy, process modeling, and simulation-optimization. In Chapter 3, we propose a personnel evaluation simulation model that represents the current evaluation and promotion systems in the United States Army and allows us to quantify the degree to which the system contributes to a suboptimal outcome of personnel evaluation. Chapter 4 introduces simulation-optimization techniques and describes their application to performance appraisal systems. These simulation-optimization techniques include ranking and selection in order to determine ordinal rankings of input parameter configurations. Chapter 5 describes areas of future research. These areas include model adjustments that provide a deeper understanding of the effect of organization structure, human behavior, and policy. Chapter 6 provides conclusions and discussions.

CHAPTER 2

LITERATURE REVIEW

2.1 Military Manpower Modeling and Simulation

> *"Operations Research is a scientific method of providing executive departments with a quantitative basis for decisions regarding operations under their control."*
>
> — Charles Goodeve, "Operational Research" in *Nature*

The field of operations research is inherently interdisciplinary with its origins deeply rooted in improving the efficiency and effectiveness of military operations. One of the earliest examples of interdisciplinary teams occurred in 1940 when physicist Patrick Blackett established a team of mathematicians, physiologists, physicists, astrophysicists, surveyors, and military officers to study techniques for improving the use of radar in anti-aircraft gunnery (Gass and Assad (2004) and Budiansky (2013)). This team, referred to as "Blackett's Circus," was later credited

with defeating the German U-Boat fleet by influencing military strategy through the application of mathematical modeling. An important aspect of operations research in World War II was that these interdisciplinary teams did not recommend new tools or weapons of warfare. Rather, they used science to recommend improved application of existing weapon systems in order to gain a competitive advantage (Morse and Kimball (1951) and Budiansky (2013)). While the field of operations research has evolved since its inception during World War II, Morse and Kimball (1951) noted that the general techniques of operations research can aid executive decisions in both industrial and governmental settings.

The study of military modeling and simulation has played a prominent role in the field of operations research. One of the most influential fields of defense modeling and simulation is military manpower modeling. Bartholomew and Forbes (1979) succinctly describe manpower modeling as an attempt to match the supply of people with the jobs available for them. Military personnel systems have long been the subject of manpower modeling due to their size relative to civilian organizations, detailed recruitment and attrition data, and clearly defined occupational specialties. Military manpower systems often provide large enough samples to draw meaningful conclusions based on aggregated data. 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. Hall (2009) determined that existing literature on military manpower modeling falls under one of three main topics: dynamic programming, Markovian models,

and goal programming. While this list is not exhaustive, it shows that the use of simulation in understanding and modeling complex manpower systems is an area that has received very little attention. Wang (2005) classified operations research techniques applied in manpower planning into four branches: optimization models, Markov chain models, computer simulation models, and supply chain management through system dynamics. Wang (2005) discussed simulation as one of the four operations research techniques applied to workforce planning, however, he was critical of simulation models, noting that they are good at answering "what happens if?" questions, but do not provide advice on the "best" workforce planning policies. In the subsequent sections, we will describe existing literature on the topic of military manpower planning using the techniques of dynamic programming, goal programming, Markovian models, and simulation models.

2.1.1 Dynamic Programming

Dynamic programming is a technique commonly applied to problems that require sequences of interrelated decisions where exhaustive enumeration of decision combinations is infeasible or extremely time-consuming. Dynamic programming problems are structured such that an optimal solution to the main problem can be determined by finding optimal solutions of its subproblems (Lew and Mauch (2007)). The optimal value of each subproblem is determined by evaluating a recursive functional equation that can be either deterministic or stochastic. Bellman (1954), credited with introducing the concept of dynamic programming, gave the

following basic functional recursion equation for dynamic programming problems:

$$f_N(\mathbf{v}) = \max_k f_{N-1}(T_k(\mathbf{v})), N = 2, 3, ..., \tag{2.1.1}$$

where the system can be described at any time by the M-dimensional vector $\mathbf{v} = (v_1, v_2, ..., v_M)$. The function $f_N(\mathbf{v})$ is the return when the stage of the system is N, and T_k is the transition as a result of choice k. In the case of stochastic dynamic programming, the transitions have associated probabilities and $f_N(\mathbf{v})$ becomes $E[f_N(\mathbf{v})]$. This makes it possible to determine the optimal set of decisions or policies one stage at a time rather than enumerating all possible combinations of decisions. The applications of dynamic programming range anywhere from long-term investment decisions, to optimal inventory and purchasing decisions, to workforce scheduling and design.

Workforce design, or manpower planning, typically consists of the combination of occupations and number of workers over a fixed or infinite time horizon. These overlapping substructures, each with an optimal solution, lends itself well to dynamic programming. Early papers on dynamic programming often cited manpower planning as an area for application. An early example of dynamic programming used in manpower models was Dailey's (1958) research on re-enlistment rates within the United States Navy. Through his research, he determined that re-enlistment rates were a function of total size of the Navy. This insight allowed Dailey to normalize annual re-enlistment rates and provided a benchmark for comparison of the annual effectiveness of re-enlistment policy. This research also

provided a predictive model for re-enlistment rates based on force size, quantifying the human behavior associated with re-enlistment. Fisher and Morton (1967) expanded upon this topic of effectiveness of re-enlistment policy. They developed multiple models based on retention incentives of a period of multiple years in order to retain technical experts requiring a high level of investment in training. This included a Cobb-Douglass production function, measuring productivity as a function of labor and capital, to compare multiple configurations of personnel strengths and capital investments. Based on the Cobb-Douglass production function, they were able to provide a framework for human resource decision making based on an ordinal ranking of configurations.

While early studies of manpower modeling used dynamic programming to evaluate the effectiveness of policy for establishing the number and type of occupational specialties, more recent research combined dynamic programming with other analytical techniques to solve workforce planning problems. Ozdemir (2013) employed stochastic dynamic programming with an analytic hierarchy processing order for personnel selection processes. The analytical hierarchy process weighed multiple criteria associated with the hiring process while incorporating the judgment of multiple decision makers, introduced in the model as relative weights. The multistage approach to dynamic programming allowed employers to evaluate and select the best candidates at each stage of the hiring process.

McGinnis and Fernandez-Gaucherand (1994) developed a decision model for optimally scheduling U.S. Army basic training resources based on dynamic pro-

gramming. Their model examined the size of basic training units, subdivided into "companies", the length of training period, and the number of idle companies for each week of a calendar year. Given the upper bounds on each one of these decision variables, this generated 4.27×10^{20} possible states for each week. While dynamic programming significantly reduces the amount of enumeration required to find an optimal solution, their problem was too large to solve using exact methods, forcing the development of multiple heuristics. The heuristics allowed them to generate and evaluate multiple training schedules based on maximizing the *quality* of training, defined as maximizing the instructor-to-student ratio. This model enabled decision making regarding quality of the training at each stage, or week, of the system without having to gauge the uniqueness and performance of the individuals within the system. This is an example of the Bellman functional recursion equation, which is a measure of the "return" of the system at each stage.

Researchers have been able to tailor the return measurement to many different manpower planning models. Rao (1990) used a dynamic programming approach to minimize the financial costs of manpower systems. Similar to Wagner-Whitin dynamic lot-size models, costs were associated with recruitment, overstaffing, understaffing, firing, and retention. While this model did not distinguish between individuals within the system, its strength was the ability to quantify the costs associated with manpower planning decisions. This is of particular interest to an organization such as the United States Army. Zais (2014) noted that within the Army's enlisted force alone, an increase of just 1% in the number of required personnel can have an adverse budget implication in excess of $1 billion.

2.1.2 Goal Programming

Price and Piskor (1972) developed a goal programming model of the Canadian military manpower planning system. The objective of this model was to assist in determining appropriate promotion quotas for various rank levels and occupational specialties over a three-year planning horizon. Their model penalized deviations associated with the financial, manpower, and promotion constraints through the use of slack and surplus variables in the following equations:

$$\min(\mathbf{c}^{+}\mathbf{r} + \mathbf{c}^{-}\mathbf{s}) \tag{2.1.2}$$

$$A\mathbf{x} + \mathbf{r} - \mathbf{s} = \mathbf{b}. \tag{2.1.3}$$

The objective function (2.1.2) contains costs, $\mathbf{c}^{+}$ and $\mathbf{c}^{-}$, that are associated with the slack and surplus variables, $\mathbf{r}$ and $\mathbf{s}$, respectively. The matrix A in Equation (2.1.3) contains the constraint coefficients while the vector $\mathbf{b}$ contains the goals of each constraint in A. This structure allowed the authors to set promotion constraints for each rank, while allowing flexibility in the mix of occupations within that rank. To do this, the costs associated with both the slack and surplus variables of the total promotion constraint were set to a graduated scale, ensuring deviations at higher ranks are penalized more severely than deviations at lower ranks. Meanwhile, the constraints with the occupational specialties had much lower costs associated with the slack and surplus variables. This gave promotion boards the flexibility to compensate for unequal distribution of high-performing

officers across multiple occupational specialties while maintaining an aggregate number of officers close to the authorized amount, particularly at higher ranks.

Georgiou and Tsantas (2002) used goal programming to develop a model for businesses to utilize European Union incentives that fund continuing education programs for hiring and promotion programs. In their model, they examined a hierarchical organization with k classes of employees. Traditionally, organizations face a push or pull promotion system. In a pull promotion system, the company waits for a position to be vacated, then train and promote an employee to fill the vacant position. In a push system, employees are preemptively promoted to reduce or eliminate the time of vacant positions. Georgiou and Tsantas (2002) proposed a system where European Union incentives are used to train additional classes of existing employees and potential recruits who remain in a *standby* position for immediate promotion or hiring within the organization. The goal programming problem minimized the costs associated with vacant positions, premature promotions, and the training expenses required minus the European Union funding for the additional *standby* classes. This novel and effective approach to the traditional promotion problem relied on the ability to accurately identify the most suitable employees and recruits for training.

Bres et al. (1980) formulated a goal programming model that differentiated between Navy officers from multiple commissioning sources with determining specific continuation rates. Based on the unique continuation rates, their model evaluated the proportion of officers from each commissioning source through the

eleventh year of commissioned service. The authors used goal programming to minimize the weighted deviations from officer strength goals required at each of the 11 out-years, while ensuring accessions from an appropriate mix of commissioning sources. Their model aggregated the cohort year groups and used historical attrition rates, focusing on commissioning source continuation rates rather than occupational continuation rates, simultaneously taking into account annual training capacity limits and acceptable upper and lower operating limits for each commissioning source. Similarly, Bastian et al. (2015) created the objective force model (OFM), a mixed-integer linear weighted goal-programming model that estimates the optimal number of hires and promotions for the U.S. Army's medical specialists. The OFM was more detailed than the model proposed by Bres et al. (1980) in that it incorporated continuation rates of multiple occupational specialties, allowing flexibility in the occupational mix of each cohort year group. This occupational imbalance, allowing a slight surplus in certain occupations without going below a minimum threshold in each occupation, provided latitude in filling skill immaterial requirements and facilitated the retention of high-performing individuals by not dictating a strict quota for each required skill.

2.1.3 Markovian Models

Many early Markovian manpower planning models were developed by studying historical transition rates between jobs, along with hirings and firings within an organization, in order to predict future requirements. As such, many manpower

planning models aggregated the workforce into occupations, grades, or total workforce strength. Wessels and van Nunen (1976) added specificity to the traditional Markov model approach in order to capture the dynamic behavior of individual employees. Rather than transition probabilities between grades being based solely on employees current grade, they were calculated using education, experience, time in current grade, and age. Even a modest enumeration of the transition criteria resulted in 900 states. This increased specificity can have ramifications in terms of usefulness of the model. Bartholomew and Forbes (1979) noted that as the number of states increases, the stocks and flows between states becomes smaller and the transition probabilities cannot be estimated with an adequate precision. Therefore, it is imperative to balance realism with the loss of accuracy by providing a reasonable amount of aggregation.

Another example of a Markov chain model used for manpower planning was Zanakis and Maret's (1980) model formulated for an organization with over 1,000 engineers. This model provided insight into predicting future organization manpower losses and position level distribution given multiple hiring quotas and growth rates. This basic application of Markov chains allowed the organization to make long-range projections, revealing the need to modify current personnel policies. While the policy decisions were made in terms of hiring quotas and promotion rates, this work demonstrated that manpower modeling can be an effective tool in informing policy. Zanakis and Maret (1980) noted that most Markovian manpower applications cited in literature were from governmental organizations, including the military, due to their size relative to civilian counterparts. While Zanakis

and Maret's (1980) research was focused on a civilian organization, their work, combined with previous literature, aided future researchers in demonstrating the ability of Markovian models to influence military policy.

Kinstler et al. (2008) developed a Markovian model for the United States Navy Nurse Corps to determine the optimal number of new recruits along with the optimal mix of commissioning sources. The U.S. Navy faced the issue of overstaffing at lower officer ranks in order to meet personnel requirements at the higher rank levels. Since officers from different commissioning sources have different continuation rates, this imbalance can be reduced by altering the mix of hiring sources. This model showed that adjusting the mix of officers entering the service can reduce the overall imbalance by 25 percent. However, the authors acknowledged that this produced disproportionate shortages at specific senior ranks, which are unlikely to be acceptable, and that they assumed static promotion rates. Their recommendation to combat shortage at specific ranks was to hire, or re-hire, officers at the more senior ranks. The practice of lateral entry, while common in the civilian sector, is not currently used in military, but some critics have advocated for its implementation (Kane (2012)). Promotions were relatively constant near 2008, but Figure 1.2 demonstrates that the assumption of static promotion rates is not realistic in the current operating environment.

Due to the dynamic nature of most organizations, many manpower planning models incorporate Markov processes into goal programming or dynamic programming. Zais (2014) noted that military manpower modeling plays a critical role

informing military leadership on system dynamics. These dynamics are often difficult to understand and translate into policy due to perpetual change. Internally, the United States military is subject to budgetary constraints, changing authorizations, and minimal surplus in personnel. Externally, personnel levels are affected by deployment uncertainties, the stochastic nature of the United States economy's impact on retention and accessions, and unknown future requirements. Zais (2014) referred to the imbalance between meeting current requirements and maintaining a posture for future requirements as *personnel friction*. Personnel friction gives the flexibility to prepare for future requirements at the expense of inducing inefficiencies in meeting current requirements. Zais (2014) modeled the career path of the enlisted forced as a Markov Decision Process (MDP), with the soldiers' decision to stay or leave at each rank and years of service. He then incorporated retention incentives and used dynamic programming to design policy in order to reduce personnel friction in the United States Army.

Hall (2009) used a Markov Decision Process (MDP) framework to model the decision of the optimal retirement point for Army officers. This model based the decision as a combination of pay and promotion opportunities within the service, pay opportunities in the civilian sector, and a taste factor which is an individual's relative preference for military service over similar, or even more lucrative, opportunities outside of the military. Even varying individual discount rates, it was shown that the optimal retirement point is at 20 years. Notably, the only exceptions to this were the times when officers were facing the potential for promotion. In those years, it was more advantageous for an officer to remain in the

service for one more year due to the potential increase in pay outweighing the increased civilian sector potential pay. Therefore, it can be reasonably assumed that promotion potential is one of the strongest factors that dictates a rational officer's decision to remain in the service.

2.1.4 Simulation Models

Of the four predominant methods for manpower modeling, simulation has received the least amount of attention in scholarly publications. Most military modeling and simulation is based on agent-based simulation, traditionally used to model interactions in a kinetic battlefield. However, discrete-event simulation has become an effective tool in understanding complex interactions and informing decision makers on a variety of non-combat related topics. Experiments using simulation often have simple assumptions, but the consequences, or *emergent properties* of the system, may not be obvious. Axlerod (1997) stated that, "if the goal [of the simulation] is to deepen our understanding of some fundamental process, then simplicity of the assumptions is important and realistic representation of all the details of a particular setting are not." In terms of social phenomena, simulation serves numerous purposes, including (Axlerod (1997)):

- Prediction - taking complicated inputs, processed by by hypothesized mechanisms, using the consequences as predictions,
- Education - allowing users to observe principles or relationships,

- Proof - providing an existence proof, i.e., demonstrating that there exists x such that conclusion,

- Discovery - using simulation for the discovery of new relationships.

Troitzsck (1997) aggregated the purposes of social science simulation into two general categories, explanatory and prediction. Explanatory models, used to account for past observations rather than predict future observations, must be done prior to the prediction form of simulation.

Lesinski et al. (2011) used the explanatory, prediction sequential construct when modeling the flow of Army officers from their commissioning source to their first operational unit of assignment using discrete-event simulation. Their purpose was to first verify and validate an officer accessions model in order to determine the degree to which the current process supports the Army Force Generation (ARFORGEN) model, a 2006 unit readiness model that replaced the previous Cold War-era model. This model incorporated multiple effectiveness measures that gauge the degree to which the timing, capacity, and duration of initial officer training support the enterprise-level unit readiness. After validating the current model, they induced perturbations in controllable parameters to determine whether there were noticeable improvements in the measures of effectiveness. Their research led to a product that allowed Army leaders to evaluate the unique readiness effect of commissioning and initial training policy.

Policy analysis was the genesis of the research conducted by McGinnis et al. (1994) when the Army Chief of Staff directed that personnel managers from the

U.S. Army Personnel Command (PERSCOM) analyze and recommend improvements to the Army's Officer Personnel Management System (OPMS). In addition to the required *key assignments* at each rank, Title IV of the 1986 Department of Defense Reorganization Act added that in order for officers to be eligible for promotion to colonel, they must complete a joint duty assignment (JDA). A JDA is an assignment working in a multi-service or multi-national command for a period of two to three and a half years. Simulation was deemed the most efficient approach to determine the feasibility of the JDA requirement since a single-period integer programming led to approximately 3.9 million integer variables, or 117 million integer variables for a 30-year model. Due to the challenges with dimensionality, discrete event simulation was a preferable alternative to traditional programming techniques. Historical point estimates and probability distributions were used for each stage of officer professional develop and promotion in order to model the impacts of Title IV of the Department of Defense Reorganization Act versus current requirements. The impacts were measured in terms of officer inventory and time required to complete all of the requirements for promotion to colonel. This model quantified the impact of policy changes associated with officer professional timelines and requirements.

Kwinn and Phelan (1996) used simulation as a tool to analyze the effect of policy on unit readiness and personnel stability. Specifically, they examined policy related to Permanent Change of Station (PCS) moves between the United States, Europe, and Asia. LTC James Thomas, Section Chief within PERSCOM, requested a tool that would allow his staff to conduct analysis of the effects of

changing personnel policies. Specifically, the model was used to determine the overall readiness impacts of a decrease in PCS moves and the effect on Korean unit readiness if European tour lengths were increased from 36 to 42 months. Given that the personnel system is stochastic in nature, Kwinn and Phelan (1996) stated that "personnel managers use most models as decision *support* tools as opposed to decision *analysis* tools." Therefore, the purpose of manpower simulation models is twofold: analyzing the current system, but more importantly, using the simulation output as justification for policy decisions.

2.2 Performance Appraisal Systems

Coens et al. (2000) defined performance appraisal as a practice that is a "mandated process in which, for a specified period of time, all or a group of employees' work performance, behaviors, or traits are individually rated, judged, or described by a person other than the rated employee and the results are kept by the organization." In general, performance appraisal systems are the means of which organizations assess and improve their employees' performance.

The use of performance appraisals in the United States Army can be traced back to 1813 when General Lewis Cass provided an assessment of his subordinate officers to the War Department (Banner and Cooke (1984)). These assessments were subjective evaluations in which Cass described each officer as anywhere from "a good-natured man" to "a knave despised by all." In 1914, the informal evaluation system was replaced by a formalized performance appraisal system that introduced

the original Department of the Army (DA) Form 67, an officer evaluation report used to assess officers in five domains: physical qualities, intelligence, leadership, personal qualities, and general value to the service (Wiese and Buckley (1998)). This assessment turned out to be effective in its ability to provide the basis for promotion for the nearly three decades, but in 1940 the same method was exposed as an ineffective tool when the War Department needed to promote 150 generals to command ground troops (Staugas and McQuitty (1950)). The weakness was that it assigned each officer an *efficiency score* that failed to distinguish between officers in the top half of their peers, labeling each of these officers as *superior* (Sisson (1948)).

In 1945, the Army Adjutant General's office tested a new form of evaluations known as *forced-choice* ratings on over 50,000 officers. Using forced-choice ratings, raters determine how well selected statements describe their subordinate officers. The responses were collected and compared to survey data from senior leaders on characteristics they desire in junior officers. While this technique rendered an officer's efficiency score obscure to some degree, it reduced the rater's ability to produce a desired outcome by choosing characteristic traits that were clearly good or bad. In 1947, the Army adopted forced-choice evaluations as the new DA Form 67-1 (Sisson (1948)).

Since 1947, the Army has modified the form used for officer evaluations nine times, with the most current form being the DA Form 67-10 (Department of the Army (2015)). During this time, the officer evaluation form has evolved into a

forced distribution performance appraisal where restrictions have been placed on raters concerning the number of officers who can be rated at the top of their peers. Raters are given a *box check* where they rate subordinates relative to their peers and provide a short narrative to accompany the box check. The DA Form 67-9 originally restricted raters from giving more than 49% of their subordinates *above center of mass* evaluations, commonly referred to as a rater's *profile constraint*. From 2004 to 2011, the profile constraint was removed for raters of officers in the rank of captain and below. When the profile constraint was reintroduced in 2011, George Piccirilli, Chief of the Evaluation, Selection, and Promotion Division at the U.S Army Human Resources Command noted, "We're bringing back honest feedback both for the rater and the senior rater. It goes back to rater accountability for fairly and accurately accessing their soldiers" (Lopez (2011)). The DA Form 67-10 implements a 49% profile constraint, now referred to as a *most qualified* evaluation, for raters of lieutenant colonel and below. Raters of colonels have a more detailed stratification where no more than 24% of their subordinates can be labeled as having "multi-star potential" and between 25% and 49% can receive the recommendation of "promote to brigadier general" (Department of the Army (2015)). Piccirilli stated that the implementation of a profile constraint better informs talent management by providing selection boards with the information needed to identify the best talent (Lopex (2011)).

There is evidence that supports Piccirilli's premise about rater accountability and the role evaluations play in the promotion board process. First, the absence of a forced distribution evaluation system often results in rater inflation. Bjerke

et al. (1987) found that the vast majority of evaluations written by commanding officers in the Navy stated that their subordinates were top 1% officers out of fear that anything less would undermine a junior officer's chances for promotion. Second, promotion boards value rater assessments of an officer's potential when selecting officers for promotion. Table 1.1 shows the promotion rates for active duty Army officers facing promotion to lieutenant colonel based on the number of above center of mass or most qualified evaluations received as a major for the boards conducted in 2015 and 2016. Officers receiving three or more top evaluations were promoted at a rate well over 70%, whereas officers receiving two or fewer top evaluations were promoted less than 20% of the time. On the surface, the officer evaluation system appears to have benefits that align with the purposes of talent identification and promotion potential, but the accuracy of evaluations is affected by several parameters and regulations.

Carroll and Schneier (1982) divided performance appraisal systems into three sequential functions shown in Figure 2.1. These three steps form the foundation for numerous human resource decisions such as compensation, promotion, demotion, training and assignment of personnel. The Civil Service Reform Act (1978) mandated the use of performance appraisals as a basis for rewarding, reassigning, promoting, and removing federal employees. Given the weight of performance appraisals, supervisors are required to establish performance standards that "permit the accurate evaluation of job performance on the basis of objective criteria." Performance appraisal systems have been scrutinized, critiqued, and modified for decades. A 1997 survey by Aon Consulting and the Society of Human Resource

Management found that only 5% of human resource professionals were "very satisfied" with their performance appraisal system (Imperato (1998)). One of the main challenges in any performance appraisal system is the ability to obtain an accurate evaluation.

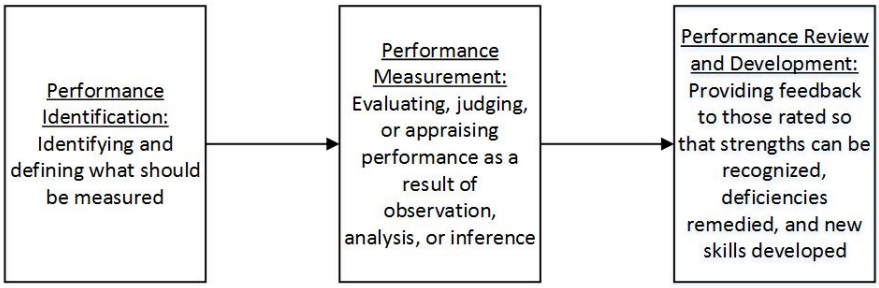


Figure 2.1: Three sequential functions of performance appraisal systems (from Carroll and Schneier, 1982)

Inaccuracy within a performance appraisal system refers to the extent that the evaluation outcome differs from the true distribution of performance levels across a group of evaluated employees (Carroll and Schneier (1982)). These inaccuracies can be the result of subjectivity of human judgment, lack of specificity in performance standards, or lack of compatibility of the performance appraisal system and the organizational structure. Inaccuracies due to human judgment can take the following forms (Coens et al. (2000), Carroll and Schneier (1982), Kozlowski et al. (1998)):

- Leniency - rater gives generous ratings in order to foster a more acquiescent workforce,
- Halo/Horn - rater forms positive (halo) or negative (horn) opinion around limited number of criteria,

- Recency - more recent accomplishments or failures are weighted more heavily,
- Severity - rater receiving a poor evaluation and believing it is due to poor performance of subordinates,
- Self-Serving - raters inflating subordinates' ratings to make themselves look better,
- Contrast/Similarity - rating employees relative to each other rather than performance standards.

While the inaccuracies due to human judgment are well-documented, errors induced by lack of specificity in performance objectives and performance appraisal system incompatibility with organizational structure are much more difficult to identify and quantify.

Physicist and mathematician W. Edwards Deming, who labeled performance appraisals one of the "Seven Deadly Diseases of Management," used the algebraic formula $IPO = X + YX$ to describe the systemic errors of performance appraisal systems, where IPO is the individual performance outcome, X is the individual contribution, and Y is the effect of the system in terms of inputs, structure, methods, tools, etc. (Elmuti (1992), Coens et al. (2000)). Deming claimed that performance appraisals attempt to quantify X while ignoring the impact of Y. There is often inherent conflict between the appraisal system and the organizational structure due to two purposes of evaluation: development and rewards (Peck (1980)). Highlighting areas for improvement makes employees look less worthy of rewards such as promotion or pay increases (Mohrman et al. (1989)). One

technique to resolve this conflict is the use of forced distribution or forced ranking evaluation systems.

Ranking techniques can be used to compare an employee's performance to those in a similar position without having identical scores that often result from non-ranked evaluations, a common outcome from systems that use Likert scales for performance criteria (Carroll and Schneier (1982)). Ranking techniques include simple ranking, paired comparisons, and forced distributions. Simple ranking is a $\{1, \ldots, n\}$ ranking of employees, paired comparisons takes the n employees and uses $[n(n-1)]/2$ pairs presented to the rater in order to determine the $\{1, \ldots, n\}$ list of ranked employees, and forced distribution typically rates the majority of employees as average with a small portion recognized as top or bottom performers. Ranking and forced distribution evaluation systems are commonly used to avoid inflation, a natural tendency given that 80% of employees view themselves as above average performers (Meyer (1980)). Ranking is often less desirable because it provides a relative, rather than an absolute, level of performance. Forced distribution systems that use extreme categories such as the top 5% or 10% can be a useful technique since most employees fall near the middle. However, this system assumes randomness of performance level distribution within an organization, which is not usually a valid assumption. Mohrman et al. (1989) also stated that forced distribution systems are better when applied to a large enough group of individuals, specifically no less than 50 employees.

Forcing distributions on smaller numbers creates inequities due to the increased

probability of an uneven distribution of high and low performers. When dealing with small samples, Lauer (2012) stated that "people observe variations that are entirely due to random chance and read into them actionable stories." This is often seen in the medical field where drawing conclusions from a small sample is widely prevalent mainly due to publication bias. Nobel Prize-winning psychologist Daniel Kahneman (2011) noted that humans seek "a view of the world around us that is simpler and more coherent than the data justify." Therefore, forced distribution performance appraisal systems must be designed to evaluate a large enough sample to obtain an accurate assessment of performance, yet not too large as to challenge the psychometric properties of the rater, commonly referred to as *span of control* (Carroll and Schneier (1982)). Psychometric properties refer to the ability of an individual to measure personal characteristics or aspects of a subordinate's job performance. Despite the challenges associated with a forced distribution evaluation system, a 2010 survey of over 750 senior level-human resource professionals found that over 30% used a specified ratings distribution (WorldatWork and Sibson Consulting (2010)).

The absence of a forced distribution has been shown to create a well-intentioned distortion of ratings in order to achieve organizational goals, referred to as *organizational context* (Kozlowski et al. (1998)). Appraisal distortion, or modifying evaluations to attain desired outcomes, has been a well-documented problem in military performance appraisal systems for nearly a century (Sisson (1948), McGregor (1957)). Prior to implementing a forced distribution performance appraisal system, the U.S. Navy saw over 97% of its officers rated in the top 1% (Bjerke

et al. (1987)). Forced distribution performance appraisal systems have become common throughout the DoD.

Three of the four services within the DoD use a form of forced distribution when documenting officer performance or promotion potential. As previously mentioned, the United States Army Evaluation Reporting System restricts raters from giving more than 49% of their subordinates "most qualified" evaluations (Department of the Army (2015)). Raters within the United States 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 United States 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 United States 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 performance appraisals. While it is clear that the absence of forced distributions has caused rating inflation, rating research has been limited in its ability to quantify the impact of system structure and *organizational context* on the effectiveness of the performance appraisal system (Kozlowski et al. (1998)).

Murphy and Cleveland (1995) posited the benefits to reducing ratings inflation

may not be worth the costs associated with achieving a more accurate performance appraisal system. Figure 2.2 illustrates the forces that can influence a rater to provide accurate of distorted ratings of subordinates. This shows that raters are likely to distort ratings when there are no rewards given to raters for providing accurate performance appraisals. Furthermore, if there are negative consequences associated with low ratings, raters are more likely to distort, or inflate, their subordinates' evaluations. Murphy and Cleveland (1995) argued that low ratings often produce consequences for the rater, consequences for the ratee, negative reactions leading to decreased productivity in the workplace, and degradation of the organization's image. Therefore, if evaluations are strictly used to determine promotions, improving the accuracy may be detrimental to an organization if it does not improve the extent to which the organization can sort their employees into the categories of promotable and non-promotable (Feldman (1986)). In addition to providing input to promotion boards on officer performance, evaluating officer performance is the foundation for identifying individuals for a broad range of operational and educational assignments in the United States Army.

2.2.1 Performance Management

Performance appraisals are typically performance *measurement* instruments that are part of a larger organizational performance *management* system (Smith and Goddard (2002)). Shutler and Storbeck (2002) stated that there are substantial opportunities for the field of operations research in analyzing performance man-

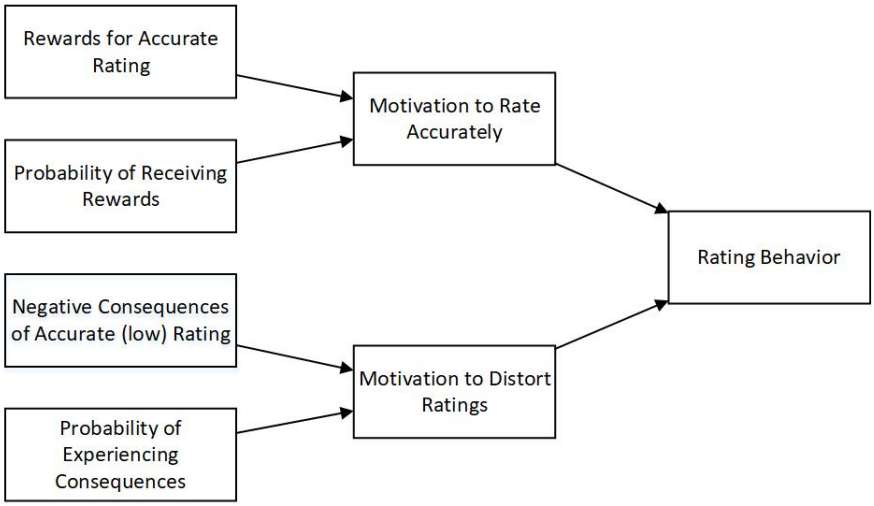


Figure 2.2: Rater motivation to provide accurate or distorted ratings (from Murphy and Cleveland, 1995)

agement since the sole focus on performance measurement can fail to achieve the overall goals of the organization. A formal definition of performance management does not exist, but performance management systems are typically comprised of four fundamental building blocks: formulation of strategy, development of performance measurement instruments, application of analytic techniques to interpret performance measurements, and development of strategies to respond to information on performance (Smith and Goddard (2002)). Armitage et al. (2012) succinctly grouped these performance management aspects into four elements: alignment, evaluation, sponsorship, and development. These four elements are further broken down into nine practices, one of which is "measuring the quality of performance appraisals."

A 2013 survey conducted by the Institute for Corporate Productivity (I4CP) revealed that only 29% of employees viewed their performance management process as fair, compared to 71% in 2006 (Stevenson et al. (2013)). Fairness is not typi-

cally an objective measure. However, employee perception of fairness is a natural extension of bias inherent to system structure and constraints. The I4CP survey also showed that only 30% of respondents said their performance management system had undergone significant changes in the past three years.

2.2.2 Talent Management

Wardynski et al. (2009) defined U.S. Army officer talent as the intersection of individual knowledge, skills, and behaviors. The authors stated that strategic-level integration of these three equally important dimensions is the essence of talent management. This definition is at odds with other traditional definitions of talent and talent management. Ulrich and Smallwood (2012) limited talent to the top 10% of individuals within an organization. Sparrow and Makramtahl (2015) stated that those with talent are a small number of elite employees who add a disproportionate amount of value to an organization. Definitions vary, but a common theme in defining talent points to the aspect of high performance (Swailes (2013)).

The recession of 2008-2009 exposed many organizations' inability to identify employees' skills, capabilities and performance levels as they attempted to downsize (Cheese (2010)). As a result, there was a renewed focus on developing data-driven talent management systems that effectively assess an employee's job performance and leadership potential. Organizational management has struggled with the defensibility of such systems due to their subjectivity, leading to a number of recent

talent management publications (Groves (2011)). Figure 2.3 shows the increased number of talent management articles published by year since 1990, with the majority published after the recession ended in June of 2009. Analyzing the retention of talent of U.S. Army officers is significant, given that nearly half of all officers leave the service within six months of the end of their initial commitment (Wardynski (2010)).

Figure 2.3: Peer-reviewed journal publications on talent management since 1990 (Source: ProQuest)

Dabkowski et al. (2010) noted that measuring officer talent is largely conceptual, but actual measurements are not necessary to assess the likely policy impacts on the retention of talent. They quantified the Army's ability to recruit and retain talent by assuming a one-dimensional distribution of talent, using historical attrition rates, and hypothesizing promotion behavior. While the normally-distributed talent score is not tied directly to a defined level of performance, it serves as a

mechanism for comparison of the impacts of different policies. Due to the complexity of the closed form solution for calculating the expected value of the r^{th} largest talent score in a sample of n observations, the authors compared the simulation results with the approximation introduced by David (1981):

$$E[X_{r:n}] = Q_r + \frac{p_r q_r}{2(n+2)} Q_r'' + \frac{p_r q_r}{(n+2)^2} \left[\frac{1}{3}(q_r - p_r) Q_r''' + \frac{1}{8} q_r p_r Q_r'''' \right] \quad (2.2.1)$$

where $p_r = r/(n+1)$, $q_r = 1 - p_r$, and Q_r is the inverse CDF of the talent score distribution evaluated at p_r. After running simulation for the no attrition scenario, they calculated a 95% confidence interval for the talent score of the least talented officer, $r = 533$, promoted to the rank of colonel. In this scenario, David's (1981) approximation fell well within the 95% simulation confidence interval computed from 1,000 replications. Thus, validating the simulation and providing a baseline model to evaluate the effect of different patterns of attrition on the retention of talent. Dabkowski et al. (2010) modeled talent as a normally-distributed variable, representing a weighted combination of multiple components.

Rather than modeling talent as a single, normally-distributed variable, Dabkowski et al. (2011) followed up their initial work by redefining officer talent as a combination of bivariate normal distribution of operational and non-operational talent. This construct allowed the authors to analyze the pool of officers and their ability to meet Army requirements at each rank as the requirements shift from an operational majority at lower ranks to a non-operational majority at higher ranks. While the authors acknowledged that talent can be developed over time, the bi-

variate normal distribution of talent was treated as *raw* talent, and therefore static over an officer's career. Howe et al. (1998) also determined that a significant part of talent is innate, but acknowledged talent could be developed through opportunities, training, and practice.

The implementation of current retention incentives suggests that the Army believes an officer's performance level talent can be identified early in the accession process. The Army has developed retention incentives targeting the commissioning sources with the most stringent screening requirements. Officers commissioning from the United States Military Academy (USMA), along with three and four-year scholarship Reserve Officer Training Corps (ROTC) officers, were given the option to attend graduate schools in return for extending their initial commitment by three years. Within ROTC, cadets are typically offered two, three, or four-year scholarships based on their qualifications, with three and four-year scholarships being the most competitive. Wardynski et al. (2010) showed in Figure 2.4 that this pre-commissioning assessment is a strong predictor of performance 20 years later in an officer's career. The goal of the graduate school incentive was to increase the number of officers from traditionally high-performing commissioning sources to continue service beyond their initial obligation, typically five years.

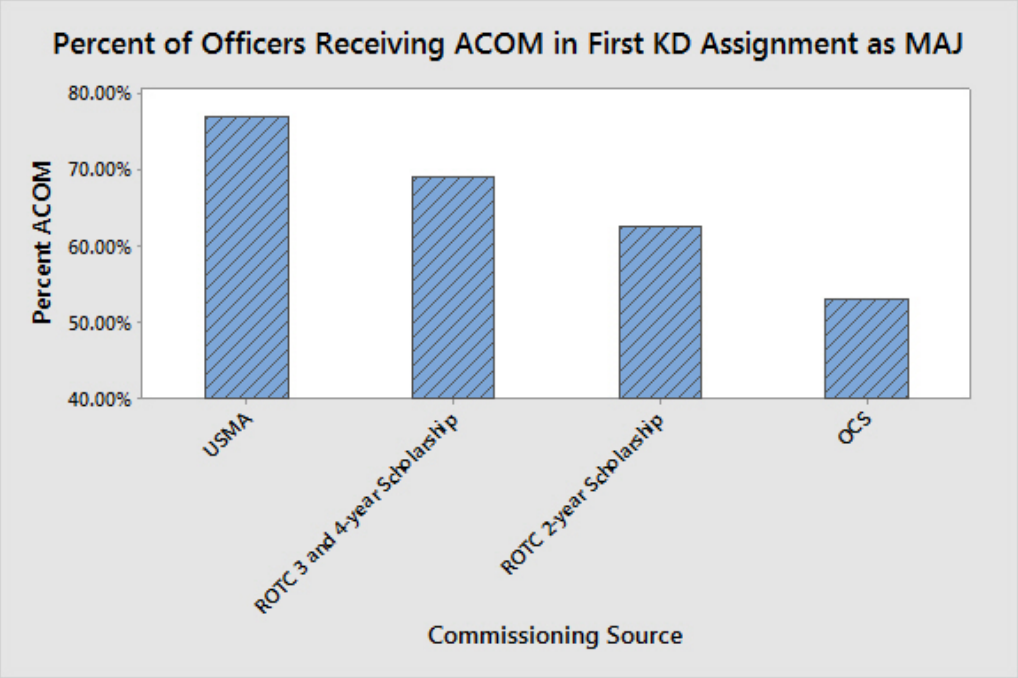


Figure 2.4: Percent of majors receiving ACOM in first key development (KD) evaluation, fiscal years 2003-2007 (Wardynski et al., 2010)

2.3 Military Policy

Military policy that governs personnel evaluation and promotion practices comes from two main sources. The National Defense Authorization Act (NDAA) dictates the number of servicemembers in each branch of the service while Army Regulation 623-3: Evaluation Reporting System provides regulatory guidance for anyone conducting an evaluation on a servicemember (114th Congress (2015), Department of the Army (2015)). As part of the overall $515 billion DoD budget, the NDAA for fiscal year 2016 authorized an active duty Army strength level of 475,000, the largest of the four services. This strength level reflected a decrease over fiscal year 2015 authorizations, in accordance with the DoD's efforts to reduce its headquarters personnel by 20%. The Army used a host of techniques to

achieve this reduction, including: selective early retirement boards (SERB), officer separation boards (OSB), and decreased promotion rates in accordance with the Defense Officer Promotion Management Act (DOPMA). Central to all of these boards was the analysis of each officer's evaluation reports.

Army Regulation 623-3: Evaluation Reporting System prescribes policy for the Army's evaluation reports that focus on the assessment of an officer's performance and potential (Department of the Army (2015)). The primary function of the evaluation reporting system is to provide information to the Department of the Army Headquarters for use in making personnel management decisions. Officers are evaluated relative to their peers using a forced distribution rating system, a practice only seen for uniformed federal employees (Office of the Under Secretary of Defense for Personnel and Readiness, (2016)). Raters are prohibited from giving more than 49% or their subordinates ACOM evaluations, with the intent that the performance of the officers receiving an ACOM evaluation exceeds the majority of the officers within the rater's population. While the maximum percent of officers receiving cannot exceed 49%, raters are encouraged to maintain a "cushion" of available ACOM evaluations to properly account for changing rated populations, commonly referred to as *rating pools* (Department of the Army (2015)).

The November 2015 revision of Army Regulation 623-3 introduced the concept of *pooling*. Pooling is defined as "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 word “pooling” was used 11 times, noting that pooling runs counter to the intent of the evaluation system and erodes soldiers’ confidence in the fairness and impartiality of their leaders. The perception of fairness is constantly challenged due to the subjectivity of ratings and the multi-criteria amalgam used in the ranking of subordinates.

Chang et al. (2007) developed a decision support system for military performance appraisal systems, treating rater evaluations as fuzzy sets. This technique is useful for ranking personnel given a multiple criteria decision making process. With the U.S. Army officer evaluation system, fuzzy sets can help explain a rater’s thought process when evaluating subordinates, much like the intersection of talent dimensions used by Dabkowski et al. (2010), but does not account for regulatory constraints placed on subjective forced distribution evaluation blocking.

2.4 Process Modeling

Process modeling is a technique commonly found in decision support literature with a heavy emphasis on business process modeling. Holt (2009) defined business process modeling as “any process modeling exercise that is performed in order to enhance the overall operation of a business.” Hangos and Cameron (2001) proposed a seven-step process modeling procedure shown in Figure 2.5. The seven-step process is recursive in nature, indicating that practitioners will regularly repeat steps in order to obtain a sufficiently refined model. Despite *a priori* knowledge of the structure of the system, the recursive procedure is necessary due to a lack of infor-

mation on the underlying mechanisms within the system that can only be gleaned through measurement data using the technique known as process identification (Hangos and Cameron (2001)).

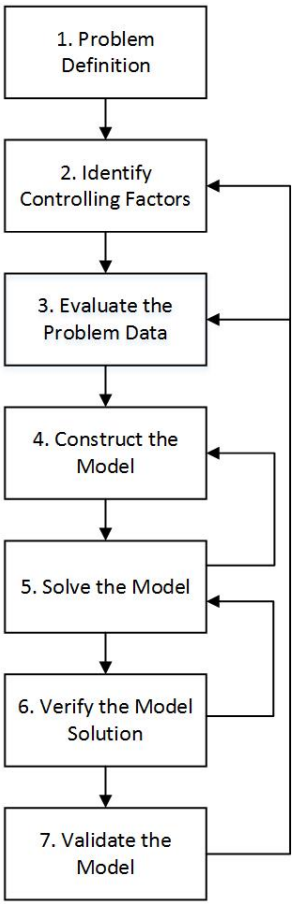


Figure 2.5: Seven step process modeling procedure (Hangos and Cameron, 2001)

System identification is used for model parameter and structure estimation and consists of the following four steps (Hangos and Cameron (2001), Zhu (2001)):

- Identification tests or experiments,
- Model order/structure selection,
- Parameter estimation,

- Model validation.

The identification tests are a structured design of experiments to obtain input-output data. The model order, or structure, refers to a set of candidate models that pertain to the particular system. In order to find the best model within the set of candidates, the user creates a cost function which is typically a sum of squared residuals of the response. The parameters used to minimize the user defined cost function, such as the function used by Ikonen and Najim (2002) in Equation (2.4.1), provide the best estimates for model validation.

$$J(\boldsymbol{\theta}) = \frac{1}{K}\sum_{k=1}^{K} \alpha_k [y(k) - \boldsymbol{\theta}^T \boldsymbol{\varphi}(k)]^2. \tag{2.4.1}$$

The cost function $J(\boldsymbol{\theta})$ assigns α_k weights to the squared differences between K observed outputs, $y(k)$, and the model predictions, $\boldsymbol{\theta}^T\boldsymbol{\varphi}(k)$. The objective is to find the parameters $\boldsymbol{\theta}$ that minimize the cost function J as in Equation (2.4.2):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} J(\boldsymbol{\theta}). \tag{2.4.2}$$

The final step in model validation is to determine whether the parameters, $\hat{\boldsymbol{\theta}}$, provide a sufficient model for the intended purposes. This is often a subjective assessment of the response metrics versus a predetermined threshold for the given model.

2.5 Simulation-Optimization

Before discussing simulation-optimization techniques, it is important to first define what a simulation is and briefly describe the purposes of a simulation. Axlerod (1997) defined simulation as "driving a model of a system with suitable inputs and observing the corresponding outputs." Given this definition, simulation has a host of diverse purposes:

- Prediction - taking complicated inputs, processed by hypothesized mechanisms, and using the consequences as predictions, identification tests, or experiments,
- Performance - using simulation to perform certain tasks such as speech recognition of medical diagnosis,
- Training and Education - allows users to observe principals or relationships,
- Proof - used to provide an existence proof (i.e., there exists as x such that *conclusion*),
- Discovery - using simulation to discover new relationships or principles.

Experiments using simulation often have simple assumptions, but the consequences may be inconspicuous. The consequences, or large-scale effects of system dynamics, are called *emergent properties* of the system. Analysis of emergent properties is done through two forms of simulation modeling, explanatory and forecasting (Troitzsch (1997)). Explanatory models are used to account for past observations

rather than predict future observations. However, a validated explanatory model is a prerequisite for models used for forecasting or prediction. Analysts can manipulate input parameters and system structures in order to estimate the effect on the system output. Determining a closed-form objective function for the simulation is often an intractable problem. Therefore, the simulation is used as a function whose explicit form is unknown, but whose output can be observed for any number of input variable and parameter settings. The basic structure of simulation optimization problems is similar to traditional optimization problems (Better et al. (2008)):

$$\begin{aligned}
&\textit{Minimize} && F(\mathbf{x}) && \text{(Objective function)} && (2.5.1)\\
&\textit{Subject to} && A\mathbf{x} \leq \mathbf{b} && \text{(Input constraints)} && (2.5.2)\\
& && g_l \leq G(\mathbf{x}) \leq g_u && \text{(Output measure constraints)} && (2.5.3)\\
& && \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} && \text{(Lower and upper bounds)} && (2.5.4)
\end{aligned}$$

The explicit form of the objective function (2.5.1) is unknown. Simulation problems can also contain input constraints (2.5.2), output measure constraints (2.5.3), as well as lower and upper bounds on the input variables or parameters (2.5.4). Given the nature of simulations, analysts cannot use deterministic optimization techniques such as linear, integer, or mixed-integer programming. Rather, researchers have developed techniques specific to simulations, referred to as *simulation optimization*.

Carson and Maria (1997) defined simulation-optimization as "the process of

finding the best input variable values from among all possibilities without explicitly evaluating each possibility." The goal of simulation-optimization is not merely enumerating a finite number of experiments and deeming the optimal inputs the best of the selected configurations. Rather, simulation-optimization techniques dictate the sequence of experiments in order to calculate the best input factors within an acceptable tolerance, or until a procedure has reached the a maximum search time limit, as shown in Figure 2.6 (April et al. (2002)). Due to the nature of the stopping criteria used in simulation-optimization and the inability to enumerate every combination of input values, the routine cannot guarantee a global optima. However, techniques such as tabu search and scatter search help overcome the problem of algorithms getting stalling at a local optimum by searching over a wide area of the solution space (Better et al. (2008)). Using these procedures, the optimization strategy takes the output of a validated simulation model, provides feedback on the progress toward achieving an optimal input parameter setting, then adjusts the inputs as necessary to improve the output.

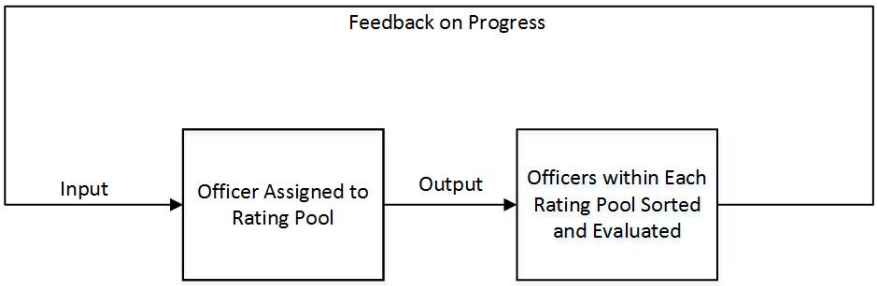


Figure 2.6: Basic logic of a simulation-optimization procedure (April et al., 2002)

Carson and Maria (1997) discussed six categories of simulation-optimization techniques shown in Figure 2.7. The following provides a brief description of each of

the six simulation-optimization categories:

- Gradient Based Search Methods - estimates the response function gradient to assess the shape of the objective function by employing deterministic mathematical programming techniques,
- Stochastic Optimization - finding a local optimum for an objective function whose values are not known, but can be estimated,
- Response Surface Methodology - fitting a series of regression models that map the input settings to the simulation output variable,
- Heuristic Models - search strategies that balance exploration with exploitation,
- A-Teams - combines strategies for multi-criteria optimization,
- Statistical Methods - sampling methods to gauge performance and compare alternatives.

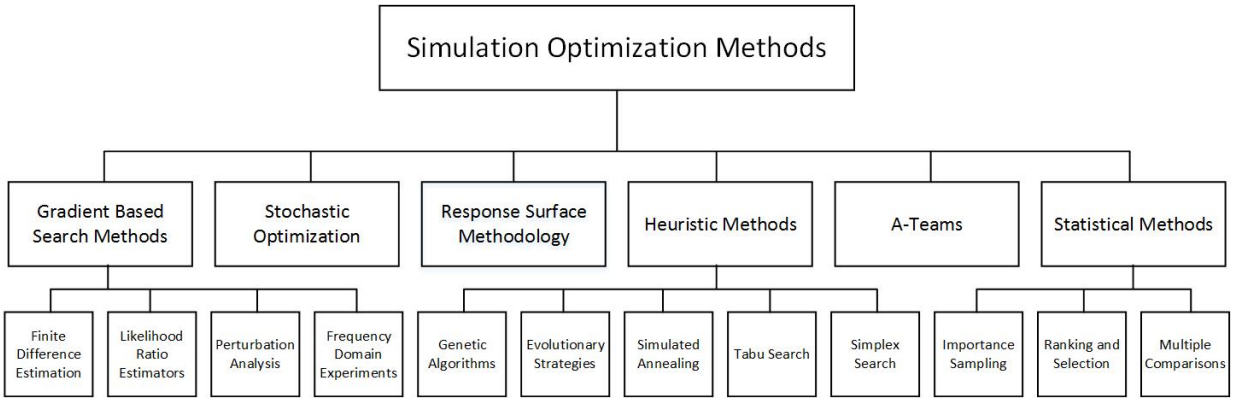


Figure 2.7: Simulation-optimization techniques (Carson and Maria, 1997)

Fu (2001) provided a similar summary of simulation-optimization categories shown in Figure 2.8. The most common of the four categories are statistical procedures and metaheuristics. Statistical procedures include response surface methodology, a metamodeling technique used in experimental design for nearly a century, and ranking and selection, a multi-stage procedure that screens and ranks a finite number of input configurations. Multiple comparison techniques are similar to ranking and selection, but provide an ordinal ranking of simulation configurations while ranking and selection techniques provide a magnitude of measurement to distinguish between configurations. Metaheuristic procedures are commonly used in commercial simulation-optimization software packages due to a combination of their difficulty to program in a general purpose programming language and their goal of finding global extrema, as opposed to local search techniques such as response surface methodology.

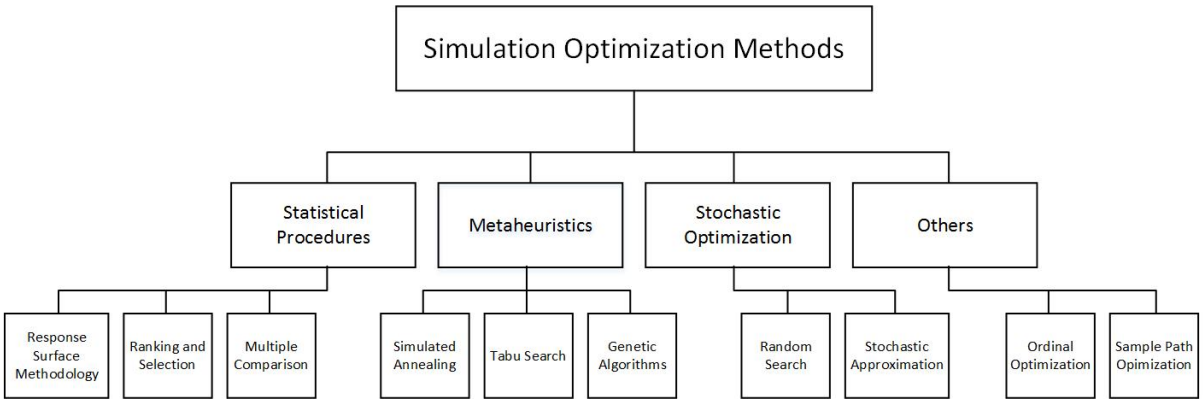


Figure 2.8: Fu's simulation-optimization techniques (Fu, 2001)

With this discussion, we now proceed to present the analytical contributions of this dissertation.

CHAPTER 3

PERSONNEL EVALUATION SIMULATION MODEL

3.1 Introduction

In this chapter, we develop and analyze a discrete event simulation model for the United States Army performance appraisal system. The model integrates system structure, system dynamics, and human behavior in order to estimate the accuracy of the performance appraisal system. In the model introduced by McGinnis et al. (1994), discrete event simulation is used to analyze the feasibility of proposed personnel policy on the U.S. Army officer corps. This model aggregates officers into cohort years groups, making no distinction between the performance level of the individual officer when evaluating policy. The effectiveness of policy is calculated by taking the percentage of officers within each cohort year group that successfully meet the proposed sequential assignment requirements. The discrete

event simulation model developed by Dabkowski et al. (2010) is a generalized model that hypothesizes the talent level of senior leaders based on multiple attrition patterns. Their premise is that if the Army can affect the retention of officers at different points in their career through retention incentives, it can increase the talent level of officers in future years. In contrast, our model quantifies the effectiveness of the performance appraisal system as a result of regulatory constraints placed on actors within the system. This model serves as the foundation for Chapter 4, in which we assess the impact of changing parameters in the system via simulation-optimization.

This chapter makes the following specific contributions. First, we present a discrete event simulation model that represents a forced distribution performance appraisal system, incorporating the structure, system dynamics, and human behavior associated with such systems. Second, we provide a means for parameter estimation to model human behavior within a complex system. The goal of this parameter estimation is to analyze human behavior and explore a method for model validation that captures the role of subordinate seniority in the evaluation process. This includes simulation experiments that map black-box functions representing human behavior to simulation outputs. The effectiveness of each behavior function is based on a multi-objective response function that is the sum of squared error measuring the difference between model outputs and historical data. Third, we quantify the theoretical error induced by the United States Army's forced distribution evaluation system. Finally, we create a response function that incorporates both the quantity and severity of the misidentifications based on historical pro-

motion data. This response function serves as the objective function, in lieu of a closed-form solution, for system analysis.

The remainder of this chapter is organized as follows. In Section 3.2, we present the performance evaluation discrete event simulation model along with the notations used for this problem. Section 3.3 describes the input analysis for both measured and estimated model inputs. In Section 3.4, we discuss model verification and validations. This includes estimating input parameters for regulatory constraints and a procedure for the bubble sort algorithm parameter estimation, simulating human behavior in the system. In Section 3.5, we conduct a preliminary output analysis and run experiments to test the impact of pool sizes and amount of time spent in each assignment on the output of the simulation. Section 3.6 discusses response function development based on the simulation results and historical data to be used for the simulation-optimization techniques used in Chapter 4.

3.2 Model Description and Notation

Advanced analytical tools can effectively capture complex system structure and dynamics, as well as human behavior and their interactions within the U.S. Army's performance appraisal system. The hypergeometric distribution provides insight into the error associated with forced distribution performance appraisal systems where officers are assigned to rating pools 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% ACOM evaluations, for a maximum of four in each 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 evenly distributed into the ten rating pools, all 40 would receive the appropriate ACOM rating. Conversely, the remaining 60 would appropriately receive an evaluation other than an 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:

$$P(X = k) = \frac{\begin{pmatrix} K \\ k \end{pmatrix} \begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}},$$

where k is the number of successes drawn in the sample size n. For a pool size of ten, $P(X = 4) = 0.264$, i.e., the probability that a rater receives and rewards exactly four of the officers deserving an ACOM evaluation is 0.264. Furthermore, $P(X > 4) = 0.361$, meaning 36.1% of the time the rater will not have enough 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.

Due to the tendency of evaluation levels being tied to the position rather than performance, a trend identified by Kane (2012), we use a data set of functional area officers. According to the Department of the Army Headquarters (2014), a functional area is a "grouping of officers by technical specialty or skills other than an arm, service, or branch that usually requires unique education, training, and experience." Kane (2012) observed that officers in *key developmental* positions identified by certain branches in which officers are required to serve as a prerequisite for promotion, often received strong evaluations as a rite of passage, while those in the queue for key developmental positions were given average performance evaluations. Using functional area data mitigates this effect because functional area officers do not have key developmental assignments. Therefore, using a subset of officers that have homogeneity of assignments provides a better representation of performance levels, absent the effect of key developmental assignments.

The notations used in our model is based on previous work by Wessels and van Nunen (1976) and Bartholomew and Forbes (1979). Wessels and van Nunen (1976) developed a qualification index, $\{q : 1, 2, \ldots, Q\}$ based on education and experience. Similarly, we assume that officers enter the system with an initial

performance percentile, Q_i. Bartholomew and Forbes (1979) proposed a model that uses a conditional probability for promotion based on the number of years in grade. They state that if it is known *a priori* that seniority affects the chance of promotion, then the main opportunity for improving the fit of a model is to choose the classes or grades with the chance for promotion being relatively constant. Since the DoD convenes promotion boards based on a cohort's time in grade, we incorporate seniority into the rater's sorting function by adjusting the performance level as a function of time, making no assumption as to whether the increased performance level is an actual improved performance level or the rater's behavior to evaluate senior officers more favorably.

The algorithm used to rank the officers within each pool, P_ℓ, is adapted from Levitin's (2003) bubble sort algorithm pseudocode, where each pass i compares the quality function output of officers j and $j+1$. Each $P_\ell[0, 1, \ldots, n_{\ell-1}]$ is the array of officers' adjusted performance percentile within each pool P_ℓ. The input is an array of orderable elements (officers) within each pool. The output is an array $P_\ell[0, \ldots, n_{\ell-1}]$ sorted in descending order:

for $i_\ell \leftarrow 0$ **to** $n_\ell - 2$ **do**

    for $j_\ell \leftarrow 0$ **to** $n_\ell - 2 - i_\ell$ **do**

        if $P_\ell[j_\ell + 1] > P_\ell[j_\ell]$ swap $P_\ell[j_\ell]$ and $P_\ell[j_\ell + 1]$.

The number of comparisons is equal to the worst case number of swaps, S_{worst}, and is a function of n_ℓ:

$$S_{worst}(n_\ell) = \frac{(n_\ell - 1)n_\ell}{2} \in \Theta(n_\ell^2). \tag{3.2.1}$$

Levitin (2003) noted that the brute-force algorithm can be inefficient for a large n_ℓ. Given that the mean $n_\ell \leq 15$ for this simulation, minimal modifications are necessary to achieve an acceptable efficiency. The only added logic is for any pass i that did not result in any j and $j+1$ swaps, the algorithm terminates. Any pass i that does not result in any j and $j+1$ swaps indicates that the elements in array P_ℓ are in descending order and any additional passes would not change the ordered array P_ℓ.

3.3 Input Analysis

Officers enter the system at a uniform rate and are assigned an attribute, Q_i, that represents the officer's initial performance percentile where $Q_i \sim \text{Uniform}(0, 1)$. An officer's initial performance percentile can be a strong indicator of future success within the Army. Dabkowski et al. (2010) showed that officers in the top quarter of their West Point class were promoted to the rank of colonel at a 54% higher rate than those officers in the bottom quarter of their West Point class. The initial performance percentile is used as one factor in the evaluation process described in detail in Section 3.4.2.

Officers are randomly assigned into rating pools for evaluation relative to their peers. The interarrival times, in days, at which officers arrive into the system is calculated by the equation:

$$\text{Interarrival Times} = \frac{\text{A}verage Pool Size}{365} \times \text{Number of Pools.}$$

The Department of the Army Secretariat sequences promotions throughout the calendar year, creating a uniform distribution of officers entering the system as a major. While Army Regulation 623-3: Evaluation Reporting System prohibits pooling and requires that raters abolish the practice, there is currently no prescribed size for rating pools (Department of the Army (2015)). Prior to the November 2015 revision of Army Regulation 623-3, pool sizes for majors varied as shown in Figure 3.1. The average pool size for the data depicted in Figure 3.1 is 15 officers.

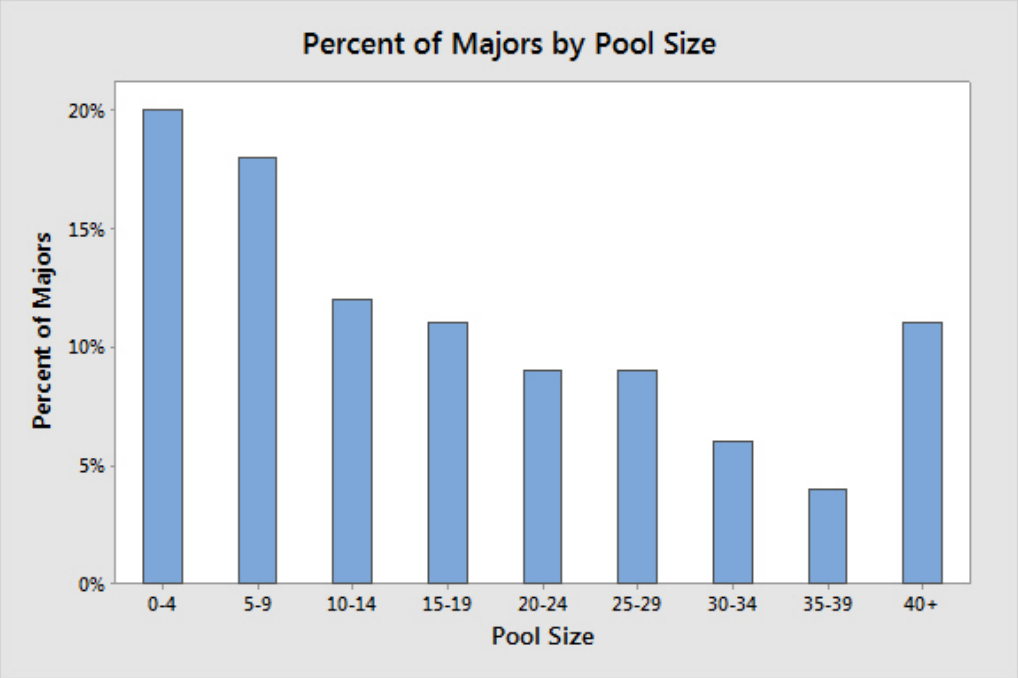


Figure 3.1: Distribution of major pool sizes (Source: U.S. Army Human Resources Command)

After each evaluation, officers either remain in the same rating pool or are assigned to a new rating pool, given that they have less than five years in the system, as depicted in Figure 1.4. The calculation of the annual probability p that an officer changes pools is shown in Equations (3.3.1)-(3.3.5). The mean

amount of time a functional area major spent in each assignment over the past five years is 16.42 months. In order for the discrete event simulation model to replicate this behavior, we must find the corresponding value of p. The structure for determining p is similar to the structure presented in Equations (2.5.1) - (2.5.4).

$$\textit{Minimize} \qquad |T(p) - 16.42| \tag{3.3.1}$$

$$\textit{Subject to} \qquad p_j = p(1-p)^{j-1}, \qquad \forall\ j = 1, 2, \ldots, 4 \tag{3.3.2}$$

$$p_5 = (1-p)^4 \tag{3.3.3}$$

$$T(p) = \sum_{j=1}^{5} 12j(p_j) \tag{3.3.4}$$

$$0 \leq p \leq 1 \tag{3.3.5}$$

In Equation (3.3.1), $T(p)$ represents E[Time in Position]. That is, the expected amount of time an officer spends in each position is a function of the probability that the officer annually changes rating pools, p. Equation (3.3.2) shows the probability that an officer stays in the same assignment 1,2,. . .,4 years. For example, the probability that an officer changes assignments after one year is $p(1-p)^{1-1} = p$. The probability that an officer remained in the same position for two years is $p(1-p)^{2-1} = (1-p)p$, which is the probability that the officer did not change rating pools after the first year times the probability the officer changed rating pools after the second year. The probability that an officer stayed in the same pool for all five years is shown in Equation (3.3.3). That is the probability the officer did not change rating pools after each of the first four years. After the fifth year, all officers exit the system. Equation (3.3.4) is the expected value for the amount of

time, in months, that an officer spends in an assignment. The probability p that minimizes Equation (3.3.1) is the optimal parameter for replicating the dynamics of the Army performance appraisal system. Table 3.1 shows the calculation of $T(p)$ for the optimal p and the corresponding probabilities that officers stay in the same assignment j years.

Table 3.1: Calculation of expected time in position for optimal $p = 0.730$.

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

3.4 Model Verification and Validation

Model verification is used to determine whether the model functions as intended. Banks (1998) stated that "the verification process involves examination of the simulation program to ensure that the operational model accurately reflects the conceptual model." Banks (1998) listed seven techniques for model verification ranging from debugging techniques to the creation of submodels. Kleijnen (2000) suggested using prior knowledge of the simulated system to determine whether changes in input values produce corresponding output values in the anticipated direction. For example, in a queueing problem, an increase in service time should produce an increase in average queue length.

Since model validation is used to determine whether that model can adequately substitute for the real system for the purpose of experimentation, simulation model

validation is both subjective and objective (Banks (2000)). Law (2015) states that "the most definitive test of a simulation model's validity is to establish that its output data closely resemble the output data that would be expected from the actual system." Since simulation can be used to model a myriad of systems, the methods for model validation are numerous. Balci (1998) compiled a list of 75 techniques for model verification and validation. However, the author acknowledged that most practitioners use a hybrid approach that is appropriate for their specific model.

3.4.1 Regulatory Constraints

According to Army Regulation 623-3: Evaluation Reporting System, raters are prohibited from giving more than 49% of their subordinates an ACOM evaluation (Department of the Army (2015)). However, an analysis of the majors in the primary zone for promotion during the years 2015 and 2016 showed that they received an ACOM evaluation 53.5% of the time. At least two factors explain this gap between the regulatory constraint of 49% and the observed value of 53.5%. First, many officers who receive only average evaluations leave the service before their primary zone for promotion. This leads to officers who have a disproportionately high number of average evaluations leaving the system. Additionally, nearly 70% of officers receive an ACOM evaluation just prior to their primary zone for promotion according to Figure 3.3, but it is common practice for officers to receive an average evaluation after the board has convened since that evaluation

will not be seen by the current board. Therefore, the profile constraint of discrete event simulation must be greater than the regulatory constraint of 49% in order to account for 53.5% of ACOM evaluations.

Determining the profile constraint that leads to 53.5% ACOM evaluations can be solved through brute force trial and error, but a more efficient method is utilizing the metaheuristics of OptQuest, incorporating tabu search and scatter search techniques (April et al. (2002)). We begin by running 50 scenarios with the profile constraint varying between 0.535 and 0.600. Given the 50 outputs from the tested scenarios, we calculate which scenario has the closest to 53.5% of ACOM evaluations using the Kim-Nelson (KN) procedure, a fully-sequential ranking and selection simulation optimization method. The profile constraint of 0.568 resulted in the minimum difference between the simulation model output and the observed value of 53.5% ACOM evaluations.

3.4.2 Sorting Function Parameter Estimation

This section examines a method for estimating this black-box function using simulation optimization. We build a discrete event simulation model and modify the sorting function used to simulate human behavior using OptQuest. Parameters from multiple functions are evaluated to determine their goodness-of-fit in replicating rater behavior. In order to evaluate the output, we use an adaptation of the cost function in Equations (2.4.1) and (2.4.2).

The simulation model, developed in Simio, follows the framework of Figure 1.4.

After officers enter the system with an initial performance percentile, they are randomly assigned into rating pools. Annually, officers are sorted and given an evaluation, X_{ij} where:

$$X_{ij} = \text{rating of officer } i \text{ in year } j, \qquad i=1,2,\ldots, n, \qquad j=1,2,\ldots,5,$$

and

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

After each evaluation, the officer changes rating pools with probability p or remains in the same rating pool with probability $1 - p$, simulating the systems dynamics of officers changing rating pools on a regular basis. After five years of collecting evaluations, the officers exit the system and their binary performance appraisal history is recorded in an output file. A truncated simulation output file is shown in Figure 3.2

A bubble sort algorithm described in Section 3.3 is used as an annually triggered event in order to rank the officers within each rating pool based on their performance percentile and seniority. Given the data trends in Figure 3.3, the proclivity for raters to award a top evaluation increases as the officers they are rating increase in seniority. Therefore, the procedure used to sort the officers uses a combination of initial performance percentile combined with a function of the time in the system. We annotate this as Q'_i, where $Q'_i(Q_i, t, \alpha)$ for Equations (3.4.1) - (3.4.3) and $Q'_i(Q_i, t, \boldsymbol{\alpha})$ for Equation (3.4.4), t is the officer's time (years)

i	Q_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	ΣX_{ik}
1	0.251878	0	0	0	0	1	1
2	0.761521	1	1	1	1	1	5
3	0.626926	0	0	1	1	1	3
4	0.051956	0	0	0	0	0	0
5	0.822873	1	1	1	1	1	5
6	0.703377	0	1	1	1	1	4
7	0.655604	0	1	1	1	1	4
8	0.907997	1	0	1	1	1	4
9	0.271632	0	0	0	1	1	2
10	0.61137	0	0	1	1	1	3
11	0.045754	0	0	0	0	0	0
12	0.240159	0	0	0	1	1	2
13	0.865431	0	1	1	1	1	4
14	0.624998	0	0	1	1	1	3
15	0.081784	0	0	0	0	1	1
16	0.402916	0	0	0	0	1	1
17	0.890514	1	1	1	1	1	5
18	0.533056	0	0	1	1	1	3
19	0.275124	0	0	0	1	1	2
20	0.267557	0	0	1	1	1	3

Figure 3.2: Sample simulation output for 20 entities

in the system, and α is an estimated parameter used to apply a weight to the officer's time in the system for Equations (3.4.1) - (3.4.3). The vector $\boldsymbol{\alpha}$ provides weights to the officer's time in the system for Equation (3.4.4). Given the rater behavior, we analyze the goodness-of-fit for the following increasing functions:

$$\text{Linear: } Q_i' = Q_i + \alpha t \tag{3.4.1}$$

$$\text{Exponential: } Q_i' = Q_i + \alpha^t \tag{3.4.2}$$

$$\text{Power: } Q_i' = Q_i + t^\alpha \tag{3.4.3}$$

$$\text{Third Degree Polynomial: } Q_i' = Q_i + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3. \tag{3.4.4}$$

Figure 3.2 shows the simulation output for a given sorting function. The analysis of each sorting function consists of its ability to replicate the actual data

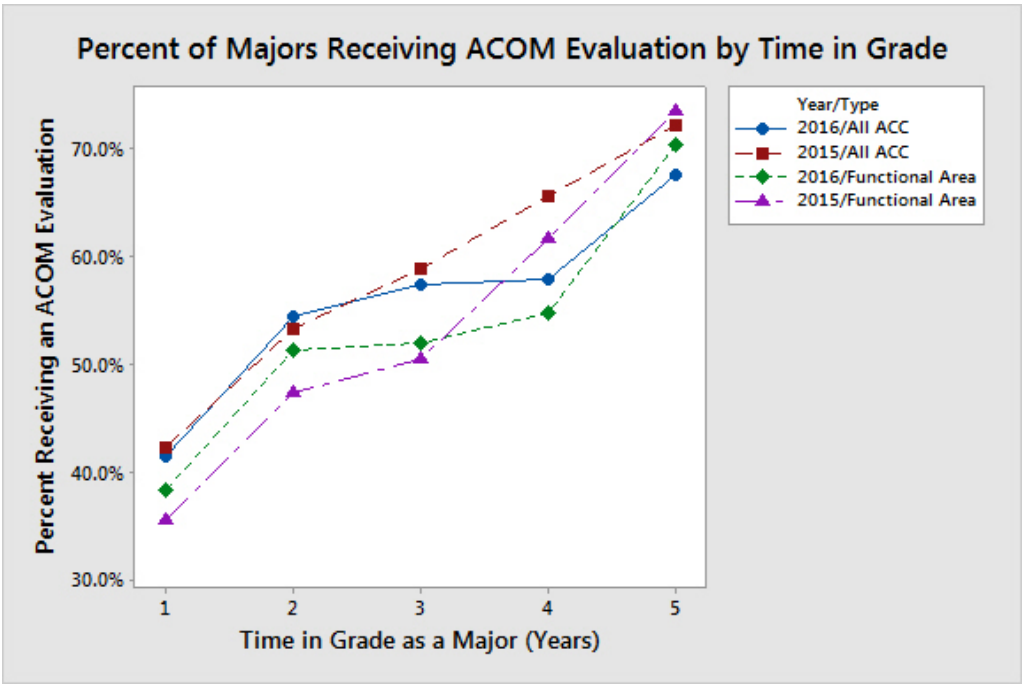


Figure 3.3: Distribution of ACOM evaluations by time in grade for U.S. Army majors in the primary zone of consideration (Source: U.S. Army Human Resources Command)

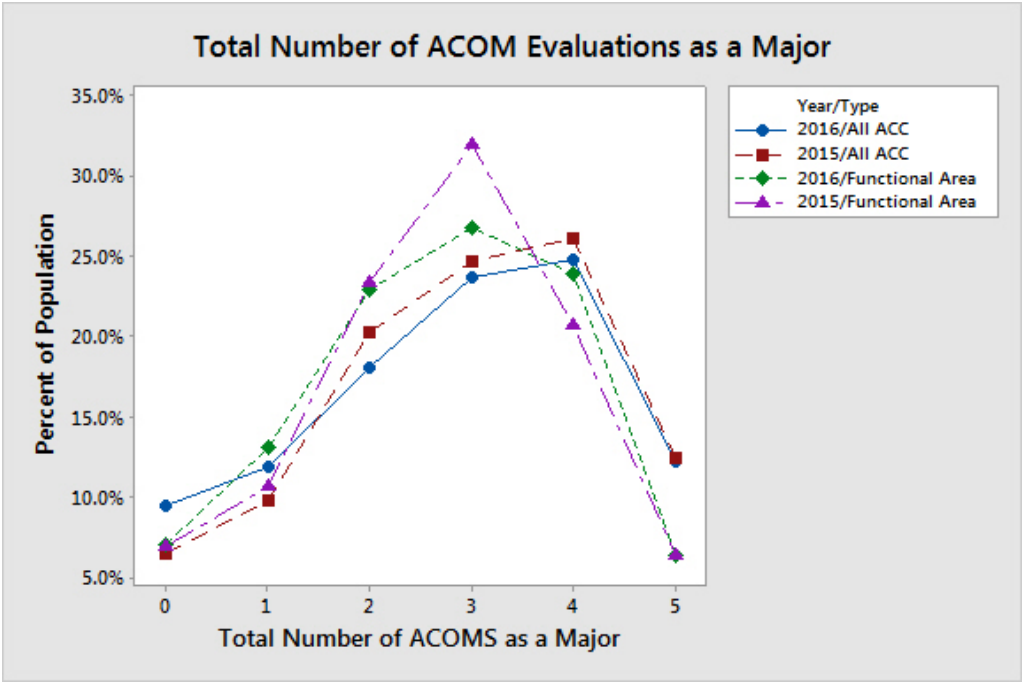


Figure 3.4: Distribution of total number of ACOM evaluations for U.S. Army majors in PZ zone of consideration (Source: U.S. Army Human Resources Command)

shown in Figures 3.3 and 3.4. This shows that a higher percentage of functional area officers receive two or three ACOM evaluations than ACC officers. Table 1.1 shows that over the past two years, officers who receive three ACOM evaluations are at least four times more likely to be promoted than officers who receive two ACOM evaluations. The data displayed in Figure 3.4 shows that functional area officers disproportionately represent the number of officers receiving two or three ACOM evaluations as compared to ACC officers, increasing their vulnerability to errors that have an impact on promotion. Before optimizing the parameters for each sorting function, it is necessary to determine a reasonable domain for α. For Equation (3.4.1), $\lim_{\alpha \to 0} Q_i' = Q_i$. An $\alpha = 0$ means that rater's determination of ranking within the rating pool is based solely off the officer's performance percentile upon entry into the system and time in the system is not a factor. Likewise, an $\alpha = 0.4$ means that the officer's time in the system is a minimum of 0.4 times as important as Q_i when $t = 1$ and a minimum of two times as important as Q_i when $t = 5$ in determining the ranking within a given rating pool. Therefore, we evaluate $0 < \alpha < 0.4$ when optimizing the output for Equation (3.4.1).

The effectiveness of Equation (3.4.2) can also be assessed using similar bounds for α. However, in Equation (3.4.2), $0 < \alpha < 1$ creates a decreasing function with respect to time in system. Furthermore, for the officer's time in the system to carry a minimum of two times the weight of Q_i when $t = 5$, $\alpha \approx 1.148$. Therefore, we limit the domain of α for Equation (3.4.2) to $0 < \alpha < 1.148$.

In order to optimize the simulation output, we use a form of the multi-objective

response function introduced by Ikonen and Najim (2002). The problem is formulated as:

$$Z_{ik} = \begin{cases} 1 & \text{if } \sum_{j=1}^{5} X_{ij} = k \qquad \forall\ k = 0, 1, \ldots, 5, \qquad i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{3.4.5}$$

where

$$k = \text{total number of top evaluations}, \qquad k = 0, 1, \ldots, 5.$$

The binary variable Z_{ik} in Equation (3.4.5) is used to identify whether each officer (i) received $0, 1, \ldots, 5$ top evaluations over the 5-year period in the system. Equation (3.4.6) measures the squared difference between the percentage of officers from the simulation with k top evaluations and A_k, where the variable A_k is the historical percentage of officers receiving k top evaluations. This squared error is calculated for each value of k in the summation:

$$T = \sum_{k=0}^{5} W_k \left(\frac{\sum_{i=1}^{n} Z_{ik}}{n} - A_k \right)^2 . \tag{3.4.6}$$

Equation (3.4.6) measures the goodness-of-fit of the simulation output compared to the data shown in Figure 3.4. The total number of top evaluations received by each officer is one measure of model accuracy. Another measure of accuracy is the timing of top evaluations each officer receives. This squared error is calculated for

each year j in the summation:

$$Y = \sum_{j=1}^{5} W_j \left(\frac{\sum_{i=1}^{n} X_{ij}}{n} - B_j \right)^2 , \tag{3.4.7}$$

where X_{ij} is the rating of officer i in year j, and B_j is the percentage of officers with a top evaluation in year j. The weights, W_k, in Equation (3.4.6) and W_j in Equation (3.4.7), allow us to control the weights of the differences between each simulation output and the actual data. This enables compensating for differences in relative error as well as the unequal number of data points in Equation (3.4.6) versus Equation (3.4.7). The value Y in Equation (3.4.7) measures the goodness-of-fit of the simulation output compared with the data shown in Figure 3.3. The measures of effectiveness provided in Equations (3.4.6) and (3.4.7) can be combined into a single weighted performance measure, $D = T + Y$. Then, the problem becomes finding the sorting function parameter value of α that minimizes the objective function D. That is, $\hat{\alpha} = \arg\min_{\alpha} D$ for Equations (3.4.1) - (3.4.3), or $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} D$ for Equation (3.4.4).

To estimate the sorting function parameters, we utilize OptQuest, the simulation optimization routine that incorporates multiple metaheuristic procedures, including tabu search and scatter search, into a single simulation optimization search procedure (April et al. (2002)). The user has the ability to modify the minimum and maximum number of replications for a specific relative error setting, along with the maximum number of scenarios. We then ran the KN procedure with an

indifference zone of 0.001 on the best subset scenarios from the OptQuest routine in order to determine optimal setting for the parameter α in each sorting function. A detailed discussion of the KN procedure can be found in Section 4.4 and in Kim and Nelson (2001). Using the Simio OptQuest add-in, 50 scenarios, with 10 replications each, took between 15 and 16 minutes to execute on an Intel® Core i5-4300U at 2.50 GHz with 8.00 GB of RAM.

For single objective parameter estimation, we performed two separate experiments to find the parameters for each sorting function that solved:

$$\hat{\alpha}_Y = \arg\min_{\alpha} \; Y \qquad \text{and} \qquad \hat{\alpha}_T = \arg\min_{\alpha} \; T \qquad \text{or}$$

$$\hat{\boldsymbol{\alpha}}_Y = \arg\min_{\boldsymbol{\alpha}} \; Y \qquad \text{and} \qquad \hat{\boldsymbol{\alpha}}_T = \arg\min_{\boldsymbol{\alpha}} \; T.$$

In Equation (3.4.7), $\mathbf{B}_j = [0.368, 0.493, 0.512, 0.582, 0.719]$, which represents the historical percent of majors that receive a top evaluation in each year in rank, j. The parameter α is evaluated in Equations (3.4.1), (3.4.2), and (3.4.3) and the minimum Y for each sorting function is shown in Figure 3.5.

Given that $\mathbf{W}_j = [1, 1, 1, 1, 1]$, each sorting function has a marked improvement over the baseline case where the performance percentile did not increase as a function of time in rank and assumed officers were evaluated strictly by this static performance level. Assuming a constant performance percentile, $Y_{baseline} = 0.0658$. Each sorting function shows a significant improvement over the baseline in Table 3.2.

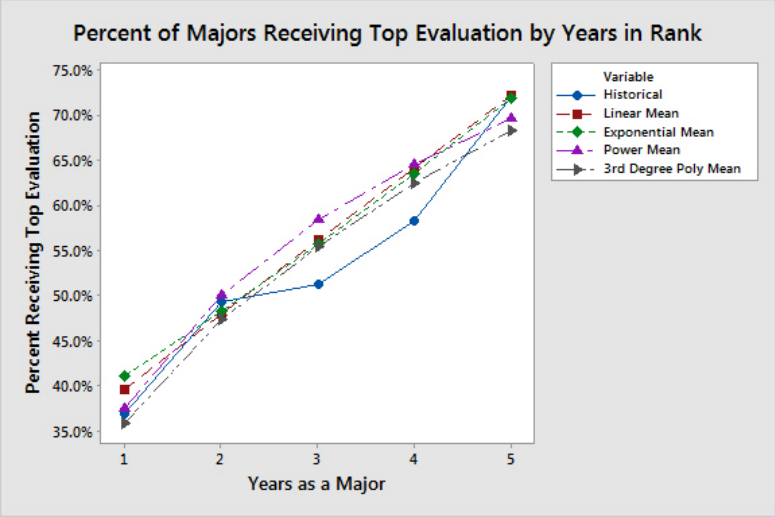


Figure 3.5: Simulation results for percent of majors receiving top evaluation by years in rank

Table 3.2: A summary of the minimum Y with sorting function parameters determined by simulation optimization.

Sorting Function	Minimum Y	Percent Improvement
Linear	0.00674	89.75%
Exponential	0.00662	89.94%
Power	0.00985	85.03%
Third Degree Polynomial	0.00506	92.31%

The parameter α is also evaluated in Equations (3.4.1), (3.4.2), and (3.4.3) and the minimum T for each sorting function is shown in Figure 3.6. In Equation (3.4.6), $\mathbf{A}_k = [0.070, 0.119, 0.231, 0.294, 0.223, 0.064]$, which represents the historical percentages of officers that receive $0, 1, \ldots, 5$ total top evaluations as a major, respectively. Given that $\mathbf{W}_k = [1, 1, 1, 1, 1, 1]$, each sorting function indicates a significant improvement over the baseline case. Assuming a constant performance percentile, $T_{baseline} = 0.298$ due to nearly 75% of officers receiving either zero or five top evaluations under the baseline approach. The results of the simulation-optimization experiments and the improvement over the baseline measurement are summarized in Table 3.3.

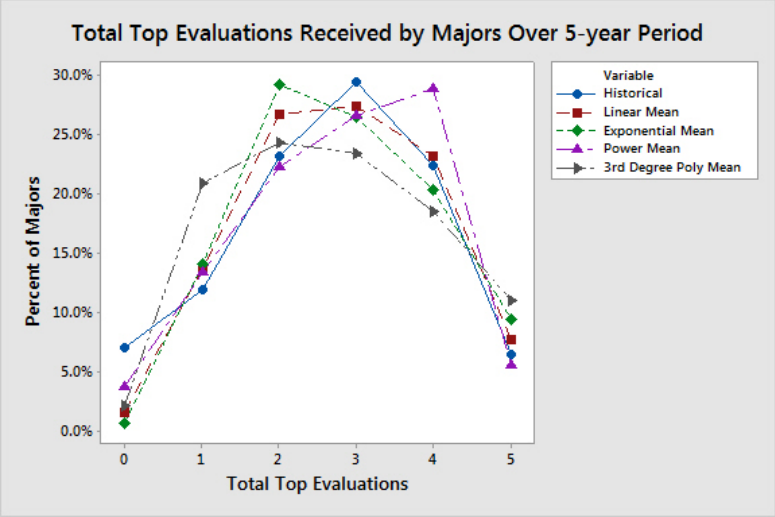


Figure 3.6: Simulation results for percentages of total top evaluations received by majors

Table 3.3: A summary of the minimum T with sorting function parameters determined by simulation-optimization.

Sorting Function	Minimum T	Percent Improvement
Linear	0.0138	95.37%
Exponential	0.0275	90.77%
Power	0.0175	94.13%
Third Degree Polynomial	0.0174	94.16%

In the single objective parameter estimation, we use separate equations for each sorting function when determining the minimum T and Y. For the multi-objective parameter estimation, we use a weighted sum of Y and T. Thus, it is necessary to determine appropriate $\mathbf{W}_j$ and $\mathbf{W}_k$ for the response function, D. Equation (3.4.7) sums the squared error between six simulation outputs and historical data, whereas Equation (3.4.6) sums the squared error between five data points and historical data. Therefore, we begin by setting each component of $\mathbf{W}_k$ to 5/6 in order to weight the outputs of T and Y equally. Finally, we factor relative error into $\mathbf{W}_k$. The mean value of the responses used in Equation (3.4.6) is 0.535, representing the average percentage of majors receiving a top evaluation in any given year.

The mean value of the responses used in Equation (3.4.7) is 0.167, representing the average percentage of majors receiving each of the six possibilities for a total number of top evaluations. We compensate for the difference in magnitudes by multiplying the initial $\mathbf{W}_k$ by 3.21 (0.535/0.167) and each component of vector $\mathbf{W}_k$ is 2.675 ($3.21 \times 5/6$). Therefore, when evaluating D, we use $\mathbf{W}_j = [1, 1, 1, 1, 1]$ and $\mathbf{W}_k = [2.675, 2.675, 2.675, 2.675, 2.675, 2.675]$. Figure 3.7 shows that minimizing D does not minimize Y or T.

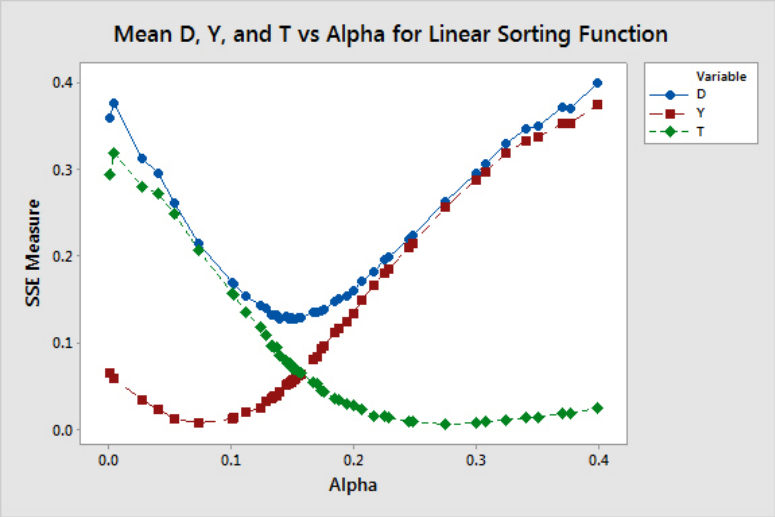


Figure 3.7: Simulation results showing relationship between D, Y, and T for linear sorting function

The efficacy of our weighted multi-objective approach is illustrated in Figures 3.8 and 3.9. The data labeled "No Time Factor" represents the baseline approach of a static performance level where $D = 0.864$. The trade-off between T and Y illustrated in Figure 3.7 results in a decreased percent improvement from the single-objective parameter estimation responses summarized in Tables 3.2 and 3.3. However, Table 3.4 shows that all three optimized sorting functions produce an output that more closely depicts the actual data over the baseline approach. This

can be seen in Figure 3.9 where the baseline approach results in over 75% of the officers receiving either zero or five top evaluations. The exceptions to this are the officers whose Q_i is near the 53.5% of the officers receiving a top evaluation. Due to the random assignment of officers into pools, officers with Q_i near 0.535 can undeservingly gain top evaluations by being assigned into a pool with a low number of strong performing officers, or conversely receive a center of mass evaluation due to an unusually high number of strong performing officers in the same pool.

Table 3.4: A summary of the minimum D with sorting function parameters determined by simulation optimization.

Sorting Function	Minimum D	Percent Improvement
Linear	0.185	78.59%
Exponential	0.208	75.93%
Power	0.181	79.05%
Third Degree Polynomial	0.183	78.94%

The experiments summarized in Table 3.4 provide insight into the weight that raters place on the experience or seniority of their subordinates. In each sorting function, a factor that is a function of time in the system, j, is added to the initial performance level Q_i for the purpose of ranking officers. We can use the function parameters to determine the added time factor for each function, at each time j. The results are summarized in Table 3.5.

Table 3.5: The weight of seniority, by year j, in the rater sorting functions.

	Year (j)				
Sorting Function	1	2	3	4	5
Linear	0.189	0.379	0.568	0.757	0.947
Exponential	1.126	1.268	1.428	1.608	1.810
Power	1.000	1.282	1.483	1.645	1.782
Third Degree Polynomial	0.188	0.361	0.521	0.670	0.810

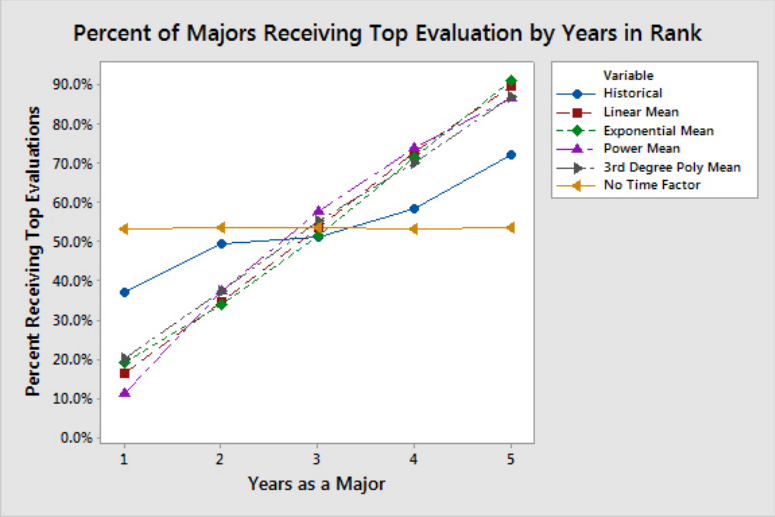


Figure 3.8: The effect on Y by minimizing weighted multi-objective response function D

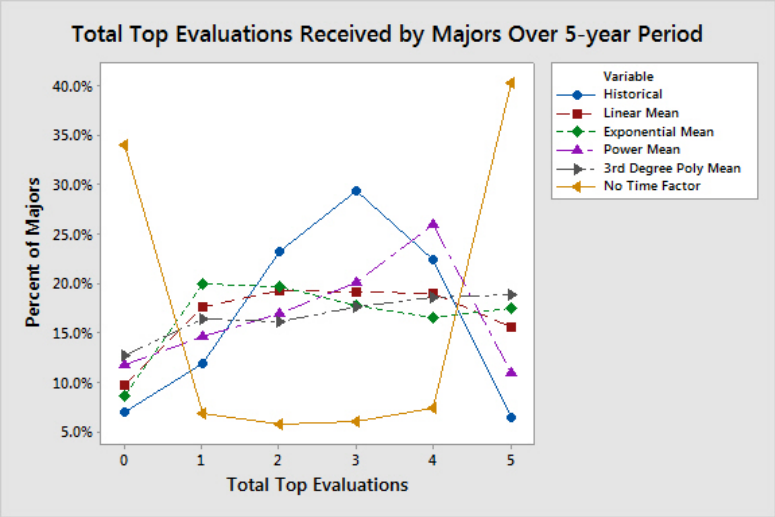


Figure 3.9: The effect on T by minimizing weighted multi-objective response function D

The results in Table 3.5 aid in model verification by showing that the optimized sorting functions produce increasing weights on time in the system. The results also provide insight as to the behavior of raters in the system and the weight they place on seniority. By measuring the difference between the year five weight and the year one weight, we can determine how raters view experienced low performers versus inexperienced high performers. The average difference between the year five

weight and year one weight for the experimental sorting functions is 0.711. An interpretation of this measure is that the highest performing officer, $Q_i = 1$, in his/her first year is viewed as the equivalent of an fifth-year officer whose $Q_i = 0.289$. The time weights displayed in Table 3.5 also provide a baseline for a discrete approach to model the rater behavior. Rather than estimating the parameters of predetermined, continuous sorting functions, we can adjust each officer's Q_i based on their years in the system using the time independent weights shown in Equation (3.4.8):

$$Q_i' = \sum_{j=1}^{5} T_{ij}(\alpha_j + \beta_j Q_i) \qquad \forall\ i = 1, 2, \ldots, n, \tag{3.4.8}$$

where

$$T_{ij} = \begin{cases} 1 & \text{if officer } i \text{ is in the } j^{\text{th}} \text{ year in the system} \quad \forall\ i = 1, \ldots, n,\ j = 1, \ldots, 5 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 < \alpha_5.$$

Table 3.6: Calculations for upper and lower bounds of α_j, $\beta_j = 1$ in Equation (3.4.8a).

	Year (j)				
Sorting Function	1	2	3	4	5
Linear	0.189	0.379	0.568	0.757	0.947
Exponential	1.126	1.268	1.428	1.608	1.810
Power	1.000	1.282	1.483	1.645	1.782
Third Degree Polynomial	0.188	0.361	0.521	0.670	0.810
Range	0.938	0.921	0.962	0.975	1.000
α_j Lower Bound	0.094	0.269	0.425	0.573	0.710
α_j Upper Bound	1.220	1.375	1.580	1.742	1.910

We establish upper and lower bounds for α_j when β is set to zero by identifying the high and low time weights at each year j in Table 3.5 and adjusting each by 10% of the range for each year. A summary of the upper and lower bounds for α_j is shown in Table 3.6. After setting upper and lower bounds for each α_j, we run the same optimization procedure used in determining the minimum D for the equations associated with the results in Table 3.4. By setting each β_j equal to one, we are able to replicate the seniority shown in Table 3.5 with Equation (3.4.8). However, unlike the equations used for the estimations in Table 3.5, Equation (3.4.8) with each β_j set to one is not constrained to continuous functions. This added flexibility of the sorting function results in an improved accuracy of the rater sorting function. The optimized parameters of Equation (3.4.8) with each β_j equal to one are shown in the row labeled Equation (3.4.8a) in Table 3.7. The optimized α_j parameters for Equation (3.4.8a) result in an 80.10% improvement over the baseline case and a 4.97% improvement over the most accurate continuous sorting function, the power sorting function.

The sorting function proposed in Equation (3.4.8) allows flexibility to the evaluation criteria used by raters. Each continuous sorting function and Equation (3.4.8a) sort officers within each rating pool according to their initial performance percentile and a function of time, with the time factor being added to the initial performance percentile. When each α_j in Equation (3.4.8) is equal to zero, as shown in the Equation (3.4.8b) row of Table 3.7, the sorting function is the product of the initial performance percentile and a scaling factor β_j. We add the

constraint

$$0 < \beta_1 < \beta_2 < \beta_3 < \beta_4 < \beta_5 < 1$$

to Equation (3.4.8) to ensure the weight placed on seniority results in an increasing function similar to Figure 3.4. Given increased range for each unknown parameter, we run the same optimization routine, but increase the number of scenarios to 500. The optimal parameter settings for each β_j are shown in the Equation (3.4.8b) row of Table 3.7. Since each β_j represents a constant multiplied with Q_i at each year j, the solution for Equation (3.4.8b) is not unique. In fact, any constant multiple of the optimized parameter vector $\boldsymbol{\beta_j}$ results in the same minimum D.

Finally, we allow Equation (3.4.8) to consist of the parameters α_j that provide an additive factor of time j, and β_j that scale the initial performance percentile according to the officer's time in the system. Given that the number of unknown parameters in Equation (3.4.8c) is double the number of unknown parameters in Equation (3.4.8a) or Equation (3.4.8b), we use an iterative process to determine the optimal parameters. We begin with the upper bounds of α_j as the values calculated for Equation (3.4.8a) and each β_j upper bound to one, and run 500 scenarios using the procedure described for the equations summarized in Table 3.4. We calculate the range of each unknown parameter by using the best subset of the 500 scenarios, and increasing each unknown parameter range by 10% as in Table 3.6. We then run 500 additional scenarios using the newly established limits on each unknown parameter. This process is repeated until the minimum D improvement is less than the predetermined indifference zone of 0.01. The

optimized parameters for all three variations of Equation (3.4.8) are shown in Table 3.7.

Table 3.7: Calculations for optimized time independent, discrete sorting function parameters α_j and β_j for use with binary variable T_{ij} and performance percentile Q_i.

	α_j					β_j				
j	1	2	3	4	5	1	2	3	4	5
Eq (3.4.8a)	0.66	0.89	1.11	1.30	1.41	1.00	1.00	1.00	1.00	1.00
Eq (3.4.8b)	0.00	0.00	0.00	0.00	0.00	0.12	0.156	0.250	0.355	0.800
Eq (3.4.8c)	0.40	0.51	0.70	0.83	0.90	0.31	0.32	0.35	0.45	0.54

All three variations of Equation (3.4.8) with optimized parameters result in improved performance over the continuous functions used in Table 3.4. Of the three versions of Equation (3.4.8), Equation (3.4.8c) results in the minimum D and will be used for the computational experiments that follow. Using the time independent weights shown in Equation (3.4.8) led to a minimum D in Equation (3.4.8c) of 0.170, which is a 6% improvement over the minimum D when using continuous functions to model rater behavior, the power function shown in Table 3.5. The performance of each sorting function is summarized in Table 3.8.

Table 3.8: A summary of the minimum D with variations of Equation (3.4.8) sorting function parameters determined by simulation optimization.

Sorting Function	Minimum D	Percent Improvement
Equation (3.4.8a)	0.172	80.10%
Equation (3.4.8b)	0.177	79.51%
Equation (3.4.8c)	0.170	80.32%

3.5 Computational Experiments

Once the input parameters have been sufficiently estimated, a series of computational experiments are constructed for the purposes of model verification and analysis of the initial results. The input parameters estimated in the previous sections include:

- Rater Profile Constraint - set at 0.568 in order to produce 53.5% of evaluations as ACOMs,
- Probability p of Changing Pools - calculated to be 0.730 in order for the $E[\text{Time in Assignment}] = 16.42$ months,
- Average Pool Size - currently 15 officers,
- Rater Sorting Function - $Q_i' = 0.397T_1 + 0.510T_2 + 0.698T_3 + 0.826T_4 + 0.904T_5 + Q_i(0.313T_1 + 0.319T_2 + 0.347T_3 + 0.445T_4 + 0.535T_5)$,
- Interarrival Rate of Officers - rate of 1.217 officers/per day produces 300 officers per year entering the performance appraisal system.

By regulation, the rater profile constraint is 49%. That is, raters are not allowed to give more than 49% of their subordinates ACOM evaluations. However, as discussed in Section 3.4.1, the majors facing promotion boards do so with 53.5% of their evaluations as ACOMs. The simulation profile constraint that corresponds to 53.5% is 0.568. The calculations for the probability p of changing pools and the average pool size are detailed in Section 3.3. The estimation of the rater sorting

function is given in Section 3.4.2. The interarrival rate of officers corresponds to 300 officers per year entering the system. This is a reasonable rate for two reasons: (1) the Army accesses and promotes approximately 300 functional area officers annually, and (2) these accessions and promotions are dispersed over a calendar year to ensure a constant flow of officers into each functional area branch. Given the discrete event simulation with the estimated input parameters, we analyze the output for model validation. Additionally, perturbations in the input parameters result in corresponding output that is used for model verification.

3.5.1 Preliminary Results and Output Analysis

The model output consists of rated officers, each with an initial performance percentile Q_i, and an evaluation vector $\mathbf{X}_i$ that consists of the evaluations received by officer i over the five-year period each officer spent in the evaluation system. We run 100 replications of the initial simulation to obtain output data on 30,000 officers. Officers spend five years in the system receiving annual evaluations. Therefore, the total number of ACOM evaluations received in the system is:

$$\sum_{j=1}^{5} X_{ij} = k \qquad \forall\ i = 1, 2, \ldots, n, \qquad \text{where } \{k : 0, 1, \ldots, 5\}. \tag{3.5.1}$$

Analyzing the distribution of Q_i for each value of Equation (3.5.1) is a critical step in evaluating the effectiveness of the system. For example, Table 3.9 shows the distribution of the simulation output as it relates to Equation (3.5.1). The

empirical distribution shown in Table 3.9 is used to determine $E[\sum_{j=1}^{5} X_{ij}]$ for each officer i. For example, if $Q_{21} = 0.411$, the expected number of k ACOM evaluations is 2 since $0.2592 < Q_{21} < 0.4412$. The lower bound, 0.2592, corresponds to the percent of officers receiving $k < 2$ ACOM evaluations and the upper bound , 0.4412, corresponds to the percent of officers receiving $k < 3$ ACOM evaluations.

Table 3.9: A summary of the percentage of officers receiving k ACOM evaluations.

k ACOM Evals	Percent of Officers ($\sum_{i=1}^{n}\sum_{j=1}^{5} X_{ij}/n$)	Cumulative Percent
0	12.16%	12.16%
1	13.76%	25.92%
2	18.19%	44.12%
3	20.85%	64.96%
4	21.33%	86.30%
5	13.70%	100.00%

The simulation output displayed in Figure 3.10 shows an increasing average Q_i for officers as the number of ACOM evaluations increase. This additionally serves as model verification since officers with a high initial performance level are more likely to receive a higher number of strong evaluations over a five-year period. The inner quartile ranges show that with the exception of the officers receiving four or five ACOM evaluations, the first quartile of each successive subset is greater than the third quartile of the previous subset. For example, the third quartile for the Q_i of officers receiving two ACOM evaluations is 0.425. The first quartile for officers receiving three ACOM evaluations is 0.468. This means that over 75% of officers receiving three ACOM evaluations have a higher Q_i than over 75% of the officers receiving two ACOM evaluations. In a perfect system, according to Table 3.9, officers with $0.2592 < Q_i < 0.4412$ would receive two ACOM evaluations. Graphically, this would be reflected by box plots for of each

successive group of officers not overlapping in Figure 3.10. However, this can be misleading since the whiskers represent outliers and do not necessarily reflect the quantity of misidentifications. The overlap in Figure 3.10 represents the magnitude of misidentifications of performance levels resulting from the simulation. The preliminary results showed that 21.00% of officers had a $Q_i > 0.4412$, meaning 21.00% of the officers receiving two ACOM evaluations should have received three ACOM evaluations. However, officer performance percentile interquartile range alone does not provide a comprehensive measure of performance appraisal system accuracy.

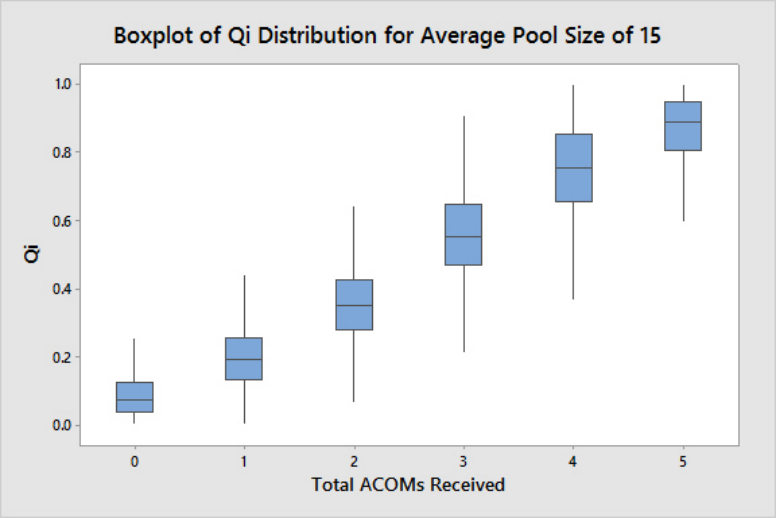


Figure 3.10: Box plot showing the distribution of Q_i for each number k of ACOM evaluations received

Tabular representations, such as classification tables, are more effective than box plots in capturing the magnitude and quantity of misidentifications within the performance appraisal simulation. The classification table shown in Table 3.10 shows the officers correctly identified and misidentified at each level k of ACOM evaluations for the current average rating pool size of 15 officers. The values in

the table refer to the percent of the total population. The columns denote the number of ACOM evaluations and the percentage of officers receiving each of the k levels. The rows represent the number of ACOMs the officers in the simulation deserved based on Table 3.9. The number of ACOM evaluations an officer deserves is based off the officer's performance percentile, Q_i, compared to the cumulative percent of officers receiving k ACOM evaluations. The diagonal of Table 3.10 displays the percentage of officers whose performance level was correctly identified and rewarded in the current performance appraisal system. The classification table also shows officers whose performance level was misidentified, along with the severity of the misidentification. For example, in the column of officers receiving two ACOM evaluations, 3.71% of the officers had a Q_i greater than 0.4412, and deserved to receive three ACOM evaluations. Similarly, 0.21% of the population received two ACOM evaluations, but had a Q_i greater than 0.6496, the equivalent of those officers deserving four ACOM evaluations. Conversely, this logic is also extended to the officers receiving more ACOM evaluations than they deserved.

3.5.2 Assessing the Effect of Pool Size

Decreasing the rating pool size results in increased interquartile ranges at each k level of ACOM evaluations. The November 2015 revision of Army Regulation 623-3: Evaluation Reporting System provides guidance for establishing appropriate pool sizes (Department of the Army (2015)). While the regulation does not state a specific pool size, it directly addresses the issue of pooling, or maintaining pool

Table 3.10: Classification table of officer misidentification in the current performance appraisal system.

		ACOM Evaluations Received					
		0	1	2	3	4	5
ACOM Evaluations Deserved	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%

sizes that are larger than the rater's ability to adequately assess the performance of the officers in each rating pool. Assuming adherence to this regulation, this change will lead to a decrease in average pool size over time. As the pool size changes, we must recalculate the profile constraint, 0.568 for the current average pool size of 15, for each new pool size experiment in order to maintain approximately 53.50% ACOM evaluations. Using the procedure outlined in Section 3.4.1, we re-calculate the profile constraint to be 0.578 for a pool size of ten and 0.649 for a pool size of five. We run the 100 replications of each simulation to obtain output data on 30,000 officers for each pool size. The results are summarized in Figure 3.11.

Decreasing the pool size has very little effect on the average Q_i at each level of total ACOMs received. However, there is a noticeable increase in the interquartile range at each ACOM level as the pool sizes decrease. This is a direct result

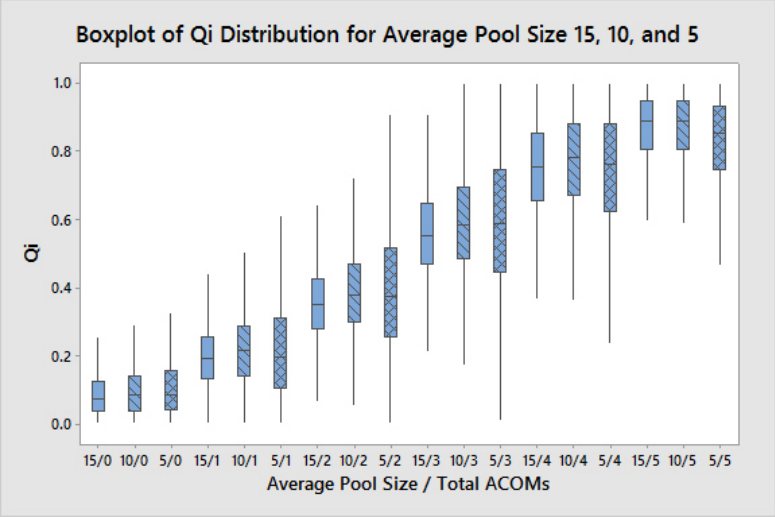


Figure 3.11: Boxplot showing the distribution of Q_i for each number k of ACOM evaluations received with varying pool sizes.

of the increased performance level variability due to small sample sizes. The interquartile range of officer performance percentile when the average pool size is five is nearly the interquartile range of officer performance percentile when the average pool size is 15. The greater variance in officer performance at each rating level for smaller pool sizes leads to an increased number and severity of potential misidentifications. The increased standard deviation and increased interquartile ranges are shown Table 3.11.

Table 3.11: The standard deviation and interquartile range of Q_i for officers receiving k ACOM evaluations for pool sizes of 15, 10, and 5.

		k					
		0	1	2	3	4	5
Avg Pool Size 15	Std Dev	0.063	0.091	0.109	0.132	0.135	0.099
	IQ Range	0.087	0.125	0.144	0.177	0.196	0.142
Avg Pool Size 10	Std Dev	0.078	0.107	0.131	0.153	0.138	0.103
	IQ Range	0.101	0.146	0.170	0.211	0.208	0.144
Avg Pool Size 5	Std Dev	0.106	0.160	0.198	0.201	0.168	0.131
	IQ Range	0.114	0.203	0.263	0.298	0.258	0.187

Increased interquartile ranges at each of the k levels leads to an increased number

of misidentifications in the performance appraisal system. The interquartile ranges at $k = 2$ and $k = 3$ nearly double from the systems with an average pool size of 15 to the systems with an average pool size of 5. The increase in interquartile range manifests into misidentifications within the system. We use the same methodology in Section 3.5.1 to sample the errors induced in the system at the $k = 2$ and $k = 3$ levels. Table 3.12 shows that 45.80% of the officers receive two ACOM evaluations of less. Therefore, any officer whose $Q_i > 0.4580$ would receive more than two ACOM evaluations in a perfect system. However, the simulation output for an average pool size of five shows that 33.57% of the officer where $k = 2$ had a $Q_i > 0.4580$. This constitutes a 59.86% increase in the number of misidentifications over the system with an average pool size of 15.

Table 3.12: A summary of the percentage of officers receiving k ACOM evaluations for an average pool size of 5.

k ACOM Evals	Percent of Officers ($\sum_{i=1}^{n}\sum_{j=1}^{5} X_{ij}/n$)	Cumulative Percent
0	9.12%	9.12%
1	16.55%	25.66%
2	20.14%	45.80%
3	23.44%	69.24%
4	21.32%	90.56%
5	9.44%	100.00%

3.5.3 Assessing the Effect of Time in Position

Varying the average amount of time each officer spends in an assignment has an effect on the accuracy of the performance appraisal system. Similar to Section 3.5.2, where we analyzed the effect of changing the pool size from 15 to five officers, we can vary the amount of time each officer spends in an assignment by adjusting

the frequency at which the officers in the simulation change rating pools. Figure 3.12 shows box plots of the performance percentile distribution of the officers receiving each level k of ACOM evaluations when the time in position is between one and five years. Increasing the 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% when the average time in position changes from one to two years. The interquartile range increases an average of 11.6% when the average time in position increases from one to five years.

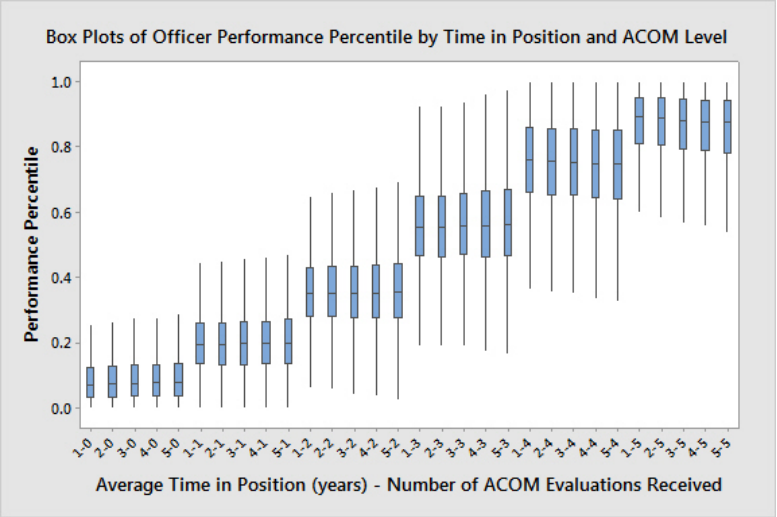


Figure 3.12: Boxplot showing the distribution of Q_i for varying time in position and k number of ACOM evaluations received.

Decreasing the amount of time and officer spends in each assignment increases the accuracy of the performance appraisal system. Table 3.13 shows the standard deviation and interquartile range of the officer performance percentiles Q_i for varying levels of average time in position. As the average amount of time officers spend in each assignment decreases from five years to one years, the standard deviation and interquartile ranges for officer performance percentiles at each level

k of ACOM evaluations received decreases. This decrease in variability indicates more accurate performance appraisal. However, the improvement in system accuracy is far less drastic than the improvements due to increases in the rating pool size. Given that moderate increases in the average time and officer spends in an assignment does not significantly affect the performance appraisal system accuracy, decreasing the frequency at which an officer moves can appear to be an effective cost cutting strategy. In 2016, the Army spent an average of nearly $19,000 per move for officers, totaling nearly $340 million for officer travel between duty locations (Deputy Assistant Secretary of the Army - Budget (2017). However, decreasing the frequency of moving an officer can serve as a hindrance to professional development. Keeping officers in assignments for longer periods of time limits the breadth of experiences critical to developing leaders of the future (Odierno (2015)).

Table 3.13: The standard deviation and interquartile range of Q_i for officers receiving k ACOM evaluations for average time in position (TIP) of 5, 4, 3, 2, and 1 years.

		k					
		0	1	2	3	4	5
Average TIP 5 Years	Std Dev	0.079	0.106	0.131	0.151	0.143	0.115
	IQ Range	0.102	0.135	0.167	0.203	0.211	0.163
Average TIP 4 Years	Std Dev	0.074	0.101	0.126	0.147	0.141	0.112
	IQ Range	0.097	0.131	0.161	0.200	0.209	0.155
Average TIP 3 Years	Std Dev	0.072	0.099	0.120	0.141	0.138	0.108
	IQ Range	0.097	0.131	0.157	0.187	0.202	0.153
Average TIP 2 Years	Std Dev	0.067	0.095	0.114	0.137	0.136	0.103
	IQ Range	0.091	0.127	0.152	0.184	0.202	0.146
Average TIP 1 Year	Std Dev	0.063	0.092	0.111	0.135	0.134	0.098
	IQ Range	0.089	0.124	0.147	0.184	0.200	0.141

3.6 Response Function Development

The response function used to evaluate the system is a combination of the number of misidentifications and the severity of the misidentifications. For example, Table 3.14 shows the percentage of officers at each k that deserved $k+1$ or $k+2$ top evaluations for an average pool size of 15.

Table 3.14: A summary of misidentified officers deserving $k+1$ or $k+2$ ACOM evaluations for an average pool size of 15.

k	Percent Deserving $k+1$	Percent Deserving $k+2$ (or more)
0	2.97%	0.21%
1	3.26%	0.15%
2	3.71%	0.21%
3	4.69%	0.43%
4	5.18%	N/A

The data shown in Table 3.15 are for an average pool size of five. It shows that at each level of k, with the exception of $k = 4$, there is a higher percentage of officers misidentified when compared to a pool size of 15. Equally as important, the officers going through the system with a pool size of five face a higher number of egregious misidentifications. The rightmost column in Table 3.15 shows that over 5% of the officers in a rating pool size of five received two or fewer ACOM evaluations than their performance percentile dictated. Furthermore, Tables 3.14 and 3.15 only show the percent of officers who deserved more ACOM evaluations than they received. To accurately measure the performance appraisal system effectiveness, we must consider the officers who received more ACOM evaluations than they deserved, as well as considering the level k at which the misidentifications occur.

Table 3.15: A summary of misidentified officers deserving $k+1$ or $k+2$ ACOM evaluations for an average pool size of 5.

k	Percent Deserving $k+1$	Percent Deserving $k+2$ (or more)
0	3.51%	0.81%
1	4.38%	1.36%
2	4.97%	1.78%
3	5.89%	1.66%
4	4.14%	N/A

The consequences associated with performance misidentifications vary across each level k. According to Table 1.1, officers who receive three or more ACOM evaluations were promoted at a rate greater than 70% for 2015 and 2016, whereas officers receiving two or fewer ACOM evaluations were promoted at a rate of less than 20%. Additionally, in terms of promotion rates, there is very little difference in promotion rates between officers who receive zero or one ACOM evaluations. Because of this, we have classified a subset of misidentifications as *critical misidentifications*. Critical misidentifications occur when officers deserved at least three ACOM evaluations, but received two or less, or officers who received three or more ACOM evaluations, but deserved two or less. The percent of critical misidentifications for a pool size of 15 is highlighted in Table 3.16.

Experiments where the average rating pool size varies show that there is an inverse relationship between the number of misidentifications and the average pool size. Table 3.17 and 3.18 are classification tables that show misidentifications and critical misidentifications for average rating pool sizes of 10 and five officers. As the rating pool sizes are decreased, there is a corresponding increase in both the misidentifications and critical misidentifications contained within the simulation output. The increased number of misidentifications is a direct result of increased

Table 3.16: Classification table of officer misidentification in the current performance appraisal system with an average rating pool size of 15 officers.

		ACOM Evaluations Received					
		0	1	2	3	4	5
ACOM Evaluations Deserved	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%
				Misidentifications			42.02%
				Critical Misidentifcations			**8.24%**

variability due to smaller sample sizes (rating pools), but the relationship between performance appraisal system accuracy and average pool size is non-linear.

Tables 3.16 and 3.17 show that decreasing the average rating pool from the current size of 15 to 10 officers results in a 19.5% increase in the number of critical misidentifications. However, comparing Tables 3.16 and 3.18 shows that the number of critical misidentifications nearly doubles. The non-linear relationship between system accuracy and average rating pool is shown in greater detail in Figure 3.13, comparing the misidentifications and critical misidentifications for average rating pool sizes ranging from five to 15 officers, in increments of two. Figure 3.13 shows that moderate decreases in the rating pool size are more likely

Table 3.17: Classification table of officer misidentification with an average rating pool size of 10 officers.

		ACOM Evaluations Received					
		0	1	2	3	4	5
ACOM Evaluations Deserved	0	9.62%	3.24%	0.20%	0.00%	0.00%	0.00%
	1	3.26%	7.45%	3.42%	**0.24%**	0.00%	0.00%
	2	0.37%	3.66%	9.54%	**4.12%**	**0.32%**	0.00%
	3	0.00%	**0.33%**	**4.38%**	10.82%	5.12%	0.61%
	4	0.00%	0.00%	**0.46%**	5.38%	10.93%	5.07%
	5	0.00%	0.00%	0.00%	0.81%	4.94%	5.71%
				Misidentifications			45.94%
				Critical Misidentifcations			**9.85%**

to be deemed acceptable by organizational leadership as opposed to drastic decreases in rating pool size that greatly decrease performance appraisal system accuracy. The effect that average rating pool size has on system accuracy is much more pronounced than the effect of average time in position.

The average amount of time officers spend in each assignment has little effect on performance appraisal system accuracy. Section 3.5.3 describes performance percentile interquartile range changes at each level k of ACOM evaluations due to changes in the average time in position. Changing the average time in position from one year to five years results in an average Q_i interquartile range increase of 2.2%. Although the effect of time in position does not appear to be significant,

Table 3.18: Classification table of officer misidentification with an average rating pool size of 5 officers.

		ACOM Evaluations Received					
		0	1	2	3	4	5
ACOM Evaluations Deserved	0	4.80%	3.50%	0.66%	0.00%	0.00%	0.00%
	1	3.51%	7.33%	4.49%	**1.03%**	0.00%	0.00%
	2	0.81%	4.38%	8.25%	**5.37%**	**1.37%**	0.00%
	3	0.00%	**1.36%**	**4.97%**	9.51%	6.50%	1.60%
	4	0.00%	0.00%	**1.78%**	5.89%	9.34%	4.74%
	5	0.00%	0.00%	0.00%	1.66%	4.14%	3.01%
				Misidentifications			57.76%
				Critical Misidentifcations			**15.88%**

Figure 3.14 shows that there is a 20.5% increase in the number of critical misidentifications when the average time in position is increased from one year (8.18%) to five years (9.86%). While the average time in position has less of an influence than average rating pool size on performance appraisal system accuracy, Figure 3.14 shows both misidentifications and critical misidentifications increase when officers stay in positions for longer periods of time.

The difference in magnitude between the misidentifications and critical misidentifications suggests that the majority of misidentifications are not egregious errors. In addition to calculating the number of critical misidentifications, giving the system an allowable error is a way to gauge the severity of misidentifications. If we

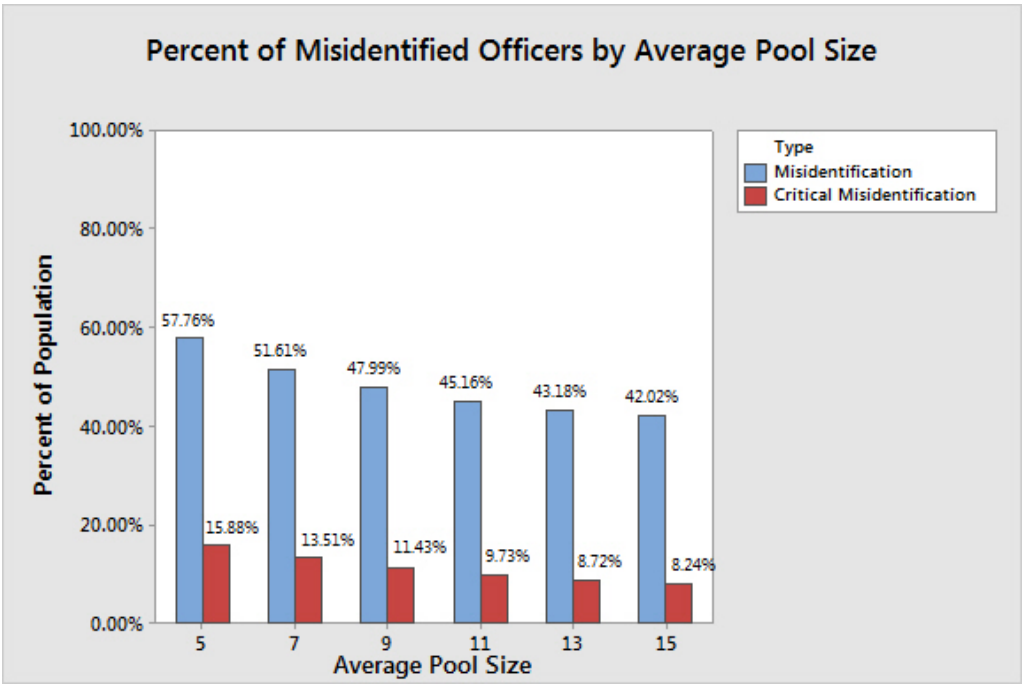


Figure 3.13: Percent of officer misidentifications and critical misidentifications when varying the average rating pool size.

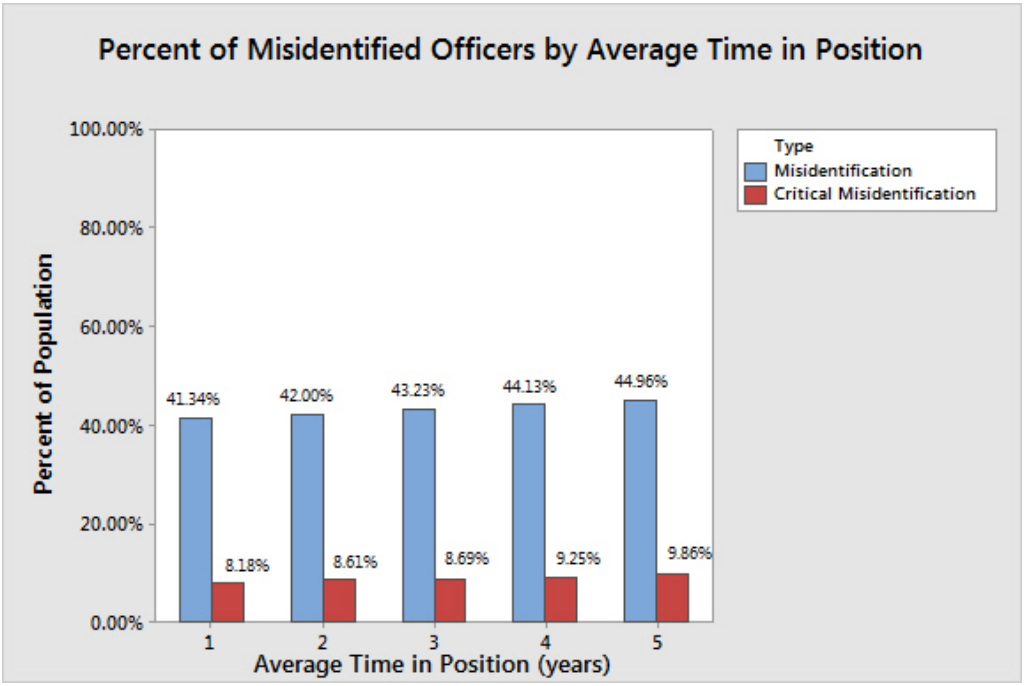


Figure 3.14: Percent of officer misidentifications and critical misidentifications when varying the average time in position pool size.

consider a 3% allowable error at each boundary Q_i cutoff, the number of misidentifications decreases significantly. Table 3.19 shows that for an average rating pool size of 15 officers, the number of misidentified officers decreases from 42.02% to 29.43% when an allowable error of 3% is applied. Additionally, the number of critical misidentifications decreases from 8.24% to 5.57%. Therefore, it is a reasonable conclusion that the misidentifications in the current performance appraisal system frequently occur when an officer's performance percentile level Q_i is near the cutoff score for each k level of ACOM evaluations.

Table 3.19: Classification table of officer misidentification in the current performance appraisal system with 3% allowable error.

		ACOM Evaluations Received					
		0	1	2	3	4	5
ACOM Evaluations Deserved	0	10.27%	1.69%	0.05%	0.00%	0.00%	0.00%
	1	1.79%	9.92%	2.10%	**0.03%**	0.00%	0.00%
	2	0.10%	2.10%	13.24%	**2.52%**	**0.16%**	0.00%
	3	0.00%	**0.05%**	**2.70%**	14.27%	3.75%	0.22%
	4	0.00%	0.00%	**0.11%**	3.72%	13.81%	4.42%
	5	0.00%	0.00%	0.00%	0.30%	3.62%	9.06%
				Misidentifications			29.43%
				Critical Misidentifcations			**5.57%**

The response function detailed in Chapter 4 combines the number and severity of misidentifications into a single response function. This response function serves

as the measure of effectiveness when evaluating system input parameter configurations. While the aforementioned, one factor at a time analysis provides insight into system behavior with perturbations to select input parameters, multiple system input parameter configurations can be evaluated using ranking and selection methods.

CHAPTER 4

SIMULATION-OPTIMIZATION

4.1 Introduction

In this chapter, we introduce and analyze simulation-optimization techniques that can be directly applied to the U.S. Army's performance appraisal system. Fu defines simulation-optimization as optimization of performance measures based on outputs from stochastic (primarily discrete-event) simulations (Fu et al. (2005)). Fu classifies simulation-optimization techniques into the four categories shown in Figure 2.8. Statistical techniques include sequential response surface methodology, ranking and selection procedures, and multiple comparison techniques. Metaheuristic techniques are generally adopted from deterministic optimization and include genetic algorithms, tabu search, and simulated annealing. Stochastic optimization includes random search techniques along with stochastic approximation. Finally, the catch-all bin of the 'Other' category in Figure 2.8 includes techniques such as ordinal optimization and sample path optimization. For solving the prob-

lem of comparing multiple configurations of performance appraisal system parameters for the purpose of making policy recommendations, we employ ranking and selection statistical procedures.

Given a finite set of system configurations, researchers have developed ranking and selection statistical techniques for discrete optimization via simulation that can be used to evaluate multiple system configurations (Rinott (1978), Goldsman and Nelson (1994), Kim and Nelson (2001), Nelson et al. (2001)). Ranking and selection methods are statistical procedures that guarantee a pre-specified probability of correct selection of the best combination of input parameters over a predetermined set of alternatives. Ranking and selection techniques are generally classified as multi-stage or fully sequential procedures. In two-stage ranking and selection procedures, k configurations are considered, and each configuration has an independent, normally-distributed mean response μ_ℓ and variance σ_ℓ^2, $\ell = 1, 2, \ldots, k$. If the mean responses are ordered such that $\mu_{[1]} \leq \ldots \leq \mu_{[k]}$, the goal of the ranking and selection procedure is to correctly select $\mu_{[k]}$, such that $\mu_{[k]} - \mu_{[k-1]} \geq \delta^*$ with the predetermined probability PC^*, where δ is defined as the indifference zone. The indifference zone represents the performance difference deemed practically significant. In a two stage procedure, an initial number of replications n_0 is run for each system configuration k. Based on μ_ℓ and σ_ℓ^2 for each system, an additional number of replications $n_b - n_0$, where $(n_b \geq n_0)$, are run to ensure that $\mu_{[k]}$ can be correctly identified for a given δ^* and PC^*. The details of finding n_b are described in Section 4.4.

Fully-sequential ranking and selection procedures reduce the overall simulation effort when compared to multi-stage ranking and selection procedures. With fully-sequential procedures, a minimal number of initial replications is run for each configuration. Statistically inferior configurations are eliminated from the candidate configurations. At each step thereafter, one replication is run for each remaining configuration and candidate configurations are removed from remaining configurations until the best candidate is found or the maximum number of replications is reached. Since inferior solutions are eliminated early in the experiment, fully-sequential procedures reduce the simulation effort required to find the best solution. However, the efficiency gained can be offset by switching between alternatives during each replication (Kim and Nelson (2001)).

4.2 Parameter Description and Optimization

Numerous factors contribute to the rating an individual receives in a forced distribution performance appraisal system. These factors include a rater's span of control (e.g., the number of subordinates being rated), the frequency at which individuals change raters, regulatory constraints pertaining to the number of top evaluations a rater can award, and the rater behavior. In this section, we describe a discrete event simulation that incorporates each of these inputs. We then apply ranking and selection simulation-optimization techniques to evaluate and optimize controllable parameters in the simulated system. Analysis and optimization of the evaluation model can provide insight for stakeholders making performance

appraisal policy adjustments or attempting to alter human behavior detrimental to the accuracy of the system.

Two types of data are used as inputs to the simulation model. Measurable inputs are parameters including average rating pool size, the frequency at which officers change rating pools, and regulations such as the profile constraint. On the other hand, estimated inputs such as performance percentile, Q_i, are rather theoretical and subjective, but based on reasonable assumptions as promotion and selection boards often use an order of merit list when determining their final recommendations. Finally, given the data in Figure 3.3, the simulation model utilizes a behavior function that accounts for each officer's performance percentile and seniority when ranked relative to their peers for the purposes of evaluations.

The simulation output consists of officers who have received five annual evaluations during their time in the simulated performance appraisal system. Based on the officers' performance percentiles and the number of top evaluations they received in the system, we can determine whether each officer's performance was correctly identified by raters in the performance appraisal system, i.e., the number of top evaluations received is commensurate with the officer's performance percentile. For instance, if an officer's performance percentile is 0.42 and the number of top evaluations received in the system over a five-year period is one; however, 39% of that officer's cohort received either zero or one top evaluation. This indicates that the officer in question should have received two top evaluations according to the performance percentile of 0.42 and the cumulative percent

of officers receiving zero or one top evaluations being less than 0.42 (39%). Therefore, we can estimate the quantity as well as severity of misidentifications in the performance appraisal system as a result of system structure, system dynamics, and behavior of the raters.

Table 4.1 shows a sample simulation output with the classifications used in an objective function for evaluating the accuracy of each performance appraisal system configuration. The diagonal elements of the classification refer to the percent of officers who correctly received the number of top evaluations corresponding to their performance percentile. Any off-diagonal elements of Table 4.1 represent misidentifications, previously separated into regular and critical misidentifications. In the objective function used for optimization, misidentifications are classified into three categories: regular misidentifications, severe misidentifications, and critical misidentifications. Regular misidentifications are instances when officers receive one more top evaluation than they deserve or one less top evaluation than they deserve. For example, 3.26% of the population received zero top evaluations, but deserved one top evaluation. Severe misidentifications are instances when officers receive at least two more top evaluations than they deserve or at least two fewer top evaluations then they deserve. An example of this case in Table 4.1 is that 0.37% of the population received zero top evaluations, but deserved two top evaluations. Critical misidentifications are displayed in bold in Table 4.1 and occur whenever the misidentification would likely have an impact on promotion. As shown in Table 1.1, critical misidentifications occur whenever an officer deserves three or more top evaluations and receives two or less top evaluations, or when

Table 4.1: An instance of percents of top evaluations officers deserved and received for an average time in position of one year, profile constraint of 49%, and a rating pool size of 10.

	Received					
Deserved	0	1	2	3	4	5
0	9.62%	3.24%	0.20%	0.00%	0.00%	0.00%
1	3.26%	7.45%	3.42%	**0.24%**	0.00%	0.00%
2	0.37%	3.66%	9.54%	**4.12%**	**0.32%**	0.00%
3	0.00%	**0.33%**	**4.38%**	10.82%	5.12%	0.61%
4	0.00%	0.00%	**0.46%**	5.38%	10.93%	5.07%
5	0.00%	0.00%	0.00%	0.81%	4.94%	5.71%

an officer deserves two or fewer top evaluations and receives at least three top evaluations.

Initial experiments showed that increasing the average rating pool size from 10 to 15 resulted in a corresponding 8.53% decrease in misidentifications. Larger rating pool samples reduce the variability of performance percentile for the officers receiving top evaluations, thereby decreasing the number of misidentifications, and providing further evidence of model validation. While the simulation results obtained by changing the rating pool size are intuitive, the size of rating pools is just one of several variables that affect the accuracy of the performance appraisal system. Prior to formulating our model, we present the following notations:

Indices and sets

$i \in I = \{1, 2, \ldots, 300\}$: Set of number of officers

$j \in J = \{1, 2, \ldots, 5\}$: Set of years spent in the system

$r \in R = \{1, 2, 3\}$: Set of sorting functions

Parameters

Q_i: Initial performance percentile for officer i

ω_{mp}: Penalty for misidentified officers

ω_{sp}: Penalty for severely misidentified officers

ω_{cp}: Penalty for critically misidentified officers

T_{ij}: Indicates if officer i is in year j in the system

α_{jr}: Rating function coefficient used for function r in year j

β_{jr}: Rating function constant used for function r in year j

Variables

P: Frequency at which officers change rating pools

C: Profile constraint for raters

D: Average rating pool size

Q'_{ir}: Sorting function r used for officer i

Outputs

X_{ij}: Evaluation for officer i in year j

MIS_i: Indicates officer i's performance percentile is misidentified

$SMIS_i$: Indicates a severe misidentification of officer i's performance percentile

$CMIS_i$: Indicates a severe misidentification of officer i's performance percentile

The accuracy of the system is determined by calculating the number of officers that fall into each of three classifications of misidentifications. Then, a negative

penalty is assigned to each category of misidentification where $\omega_{cp} < \omega_{sp} < \omega_{mp}$, reinforcing critical misidentifications as the least desirable outcome of the performance appraisal system. Therefore, we present the following model for evaluating performance appraisal system configurations:

$$Max\ Z = E\left[\omega_{mp}\sum_{i\in I} MIS_i + \omega_{sp}\sum_{i\in I} SMIS_i + \ \omega_{cp}\sum_{i\in I} CMIS_i\right] \tag{4.2.1}$$

Subject to

$$Q'_{ir} = \sum_{j\in J}(\alpha_{jr}Q_i + \beta_{jr})T_{ij} \quad \forall\ i \in I, r \in R \tag{4.2.2}$$

$$P \in [1.0, 0.48, 0.26] \tag{4.2.3}$$

$$C \in [0.39, 0.49, 0.59] \tag{4.2.4}$$

$$D \in [5, 10, 15] \tag{4.2.5}$$

$$Q'_{ir} \in [Q'_{i1}, Q'_{i2}, Q'_{i3}] \tag{4.2.6}$$

Equation (4.2.1) is the total penalty function that serves as a basis for comparison between competing system configurations. Constraint (4.2.2) is the sorting function Q'_{ir} estimated in Section 3.4.2. The piecewise function Q'_{ir} provides the appropriate amount of weight to seniority and performance percentile when compared to actual rater behavior. This function maps each performance percentile Q_i to a value Q'_{ir} that is used to rank officers within each rating pool. Figure 4.1 shows the three sorting functions listed in Constraint (4.2.6). The function Q'_{i1} is

the original function (Q'_i) estimated in Section 3.4.2, while Q'_{i2} reduces the effect of seniority by 50% for all $j \neq 3$ and Q'_{i3} increases the effect of seniority by 50% for all $j \neq 3$.

Constraint (4.2.3) varies the annual probability at which an officer changes rating pools. When $P = 1$, officers change ratings pools annually. When $P = 0.48$ and $P = 0.26$, the E[Time in Position] is 24 and 36 months, respectively. Constraint (4.2.4) restricts the rater's profile constraint for the maximum percent of top evaluations available for each rating pool. By regulation, the current profile constraint is set to 49%. Finally, Constraint (4.2.5) varies the average rating pool size. While the current average pool size is 15 officers, the November 2015 revision of Army Regulation 623-3 now requires raters to avoid pooling, which will likely reduce the size of rating pools over time (Department of the Army (2015)). Therefore, Constraint (4.2.5) defines the set of evaluated rating pools sizes at, or below, the current rating pool size.

The results in Table 4.1 are obtained for one particular configuration of the decision variables shown in Constraints (4.2.3)-(4.2.6). Different combinations of these decision variables will yield unique simulation outputs. We now describe two ranking and selection procedures and outline the solution algorithms.

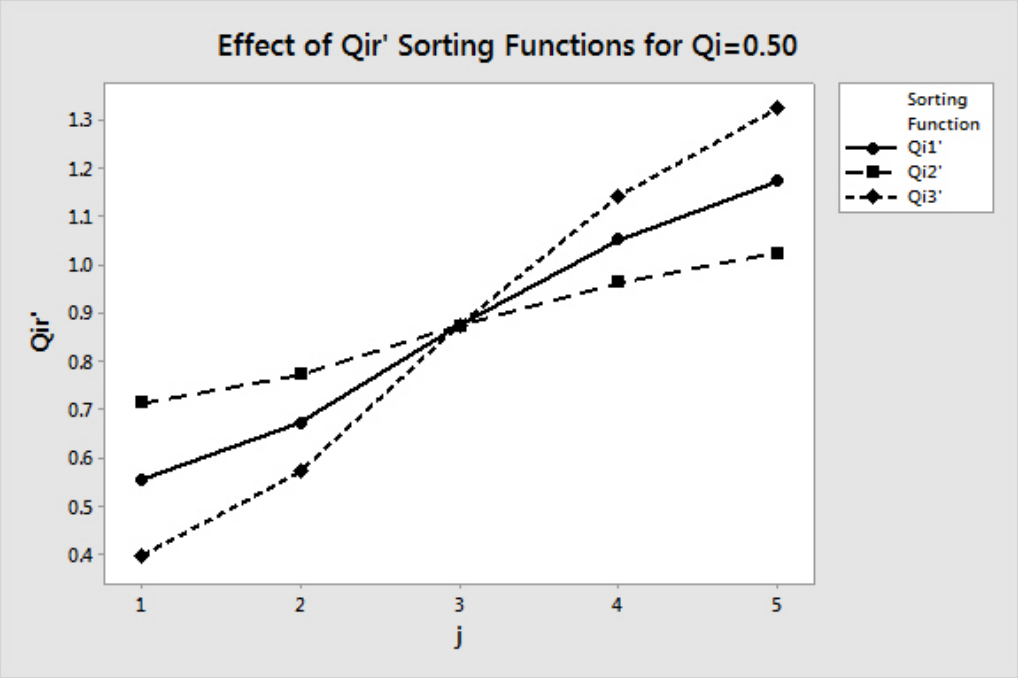


Figure 4.1: The effect of sorting function Q'_{ir} for an officer with Q_i =0.50.

4.3 Nelson, Swann, Goldsman, Song (NSGS) Procedure

A common challenge in simulation is the ability to compare the output of a large number of simulation configurations. Each configuration has a unique output, hence it is desirable to develop a technique for ranking the multiple alternatives and selecting the most preferable, or optimal, configuration. Traditional two-stage ranking and selection procedures, such as the one developed by Rinott, enable running a small number of initial replications for each system (Rinott (1978)). Then, based on the statistical properties of the mean responses, an additional number of replications is assessed and run for each system that guarantees a user-defined performance difference significance between competing systems at a specified confidence level. Rinott's (1978) procedure implements a sample-sample-select al-

gorithmic structure that is efficient for a relatively small number of competing systems. Nelson et al. (2001) adapted Rinott's procedure in order to compare a large number of competing systems by adding a screening step between the first and second sampling stages. Their sample-screen-sample-select procedure eliminates statistically inferior system configurations prior to the second sampling stage which reduces the required number of overall replications for the next stage.

In order to evaluate multiple configurations of the performance appraisal system, we use the NSGS ranking and selection procedure proposed by Nelson et al (2001). We define the notations used in the NSGS procedure, then describe the algorithmic steps as below.

n_0: initial number of replications run for each of k competing configurations

PC^*: confidence level $(1-\alpha)$ of selecting the best system, where $\frac{1}{k} < PC^* < 1$

δ^*: user-defined level of practical significance (indifference zone)

Step 1: Set $t = t_{(1-\frac{\alpha}{2})^{\frac{1}{k-1}}, n_0-1}$, where t denotes the $(1-\frac{\alpha}{2})^{\frac{1}{k-1}}$ quantile of the t-distribution with $n_0 - 1$ degrees of freedom.

Step 2: Calculate Rinott's constant h where $h = h(1-\frac{\alpha}{2}, n_0, k)$ (Bechhofer et al. (1995)).

Step 3: Sample $M_{\ell g}$, $\ell = 1, 2, \ldots, k$; $g = 1, 2, \ldots, n_0$.

Step 4: Compute the sample means and variances, $\overline{M}_\ell^{(1)}$ and S_ℓ^2, for $\ell = 1, 2, \ldots, k$.

Step 5: Let $W_{\ell g} = t\left(\frac{S_\ell^2}{n_0} + \frac{S_g^2}{n_0}\right)^{\frac{1}{2}}$ $\forall\ \ell \neq g$.

Step 6: Let the subset $L = \{\ell : 1 \leq \ell \leq k \text{ and } \overline{M}_\ell^{(1)} \geq \overline{M}_g^{(1)} - (W_{\ell g} - \delta^*),$

$\forall\ \ell \neq g\}$.

Step 7: If L contains a single configuration, it is the best solution. Otherwise,

$\forall\ \ell \in L$ compute the required replications for the second stage where

$N_b = \max\left\{n_0, \left\lceil \left(\frac{hS_\ell}{\delta^*}\right)^2 \right\rceil\right\}$.

Step 8: Take the second stage observations, $N_b - n_0$, from all remaining

configurations $\ell \in L$ and compute the overall sample means

$\overline{M}_\ell^{(2)} = \frac{1}{N_b} \sum_{g=1}^{N_b} M_{\ell g}$.

Step 9: Select the configuration $\ell \in L$ with the largest $\overline{M}_\ell^{(2)}$.

Nelson et al. (2001) make use of the probability of correctly selecting the best system such that $\mu_{[k]} - \mu_{[k-1]} \geq 1 - \alpha$ using the procedure described as above.

4.4 Kim-Nelson (KN) Procedure

Restricting the number of stages where inferior systems are removed is necessary in large part to prevent switching between a large number of system simulations. As computing efficiency improved, Kim and Nelson (2001) developed a fully-sequential procedure that takes a single observation from each competing system at each stage. The Kim and Nelson (KN) procedure enables elimination of inferior systems from contention using an indifference parameter and probability of correct selection similar to the NSGS procedure. Fully-sequential procedures implement a screening step that evaluates and eliminates system configurations

after each replication until an optimal solution is found with probability PC^*, or the experimenter has reached the maximum number of replications. We describe the algorithmic implementation of the KN procedure below.

Step 1: Let n_0 be the first stage sample size, where $n_0 \geq 2$.

Step 2: Let the subset $L = \{1, 2, \ldots, k\}$ be the set of configurations still in contention, and let $\gamma^2 = 2\eta \times (n_0 - 1)$ and $\eta = \frac{1}{2}\left[\left(\frac{2\alpha}{k-1}\right)^{-2/(n_0-1)} - 1\right]$ (Kim and Nelson (2001)).

Step 3: Sample $M_{\ell g}$, $\ell = 1, 2, \ldots, k$; $g = 1, 2, \ldots, n_0$.

Step 4: For all $\ell \neq q$, compute the sample variance

$$S^2_{\ell q} = \frac{1}{n_0 - 1}\sum_{g=1}^{n_0}(M_{\ell g} - M_{qg} - [\overline{M}_\ell(n_0) - \overline{M}_q(n_0)])^2.$$

Step 5: Let $N_{\ell q} = \left\lfloor \frac{\gamma^2 S^2_{\ell q}}{(\delta^*)^2} \right\rfloor$ and $N_\ell = \max_{q \neq \ell} N_{\ell q}$.

Step 6: If $n_0 > \max_\ell N_\ell$, the system with the largest $\overline{M}_\ell(n_0)$ is the best configuration.

Step 7: If $n_0 < \max_\ell N_\ell$, set the replication counter $b = n_0$.

Step 8: Let $L = \left\{\ell : \ell \in L^{\text{old}} \text{ and } \overline{M}_\ell(b) \geq \overline{M}_q(b) - W_{\ell q}(b), \forall\ q \in L^{\text{old}}, q \neq \ell\right\}$ where $W_{\ell q}(b) = \max\left\{0, \frac{\delta^*}{2b}\left(\frac{\gamma^2 S^2_{\ell q}}{(\delta^*)^2} - b\right)\right\}$.

Step 9: If $|L| = 1$, the configuration in L is the best solution. Otherwise, take one additional replication $M_{\ell,b+1}$ from each system $\ell \in L$, set $b = b + 1$, and return to Step 8.

4.5 Applied Simulation Optimization Results

We implemented the NSGS procedure using the performance appraisal system evaluation model. A 3^4 factorial design resulted in 81 configurations of the performance appraisal system simulation where the annual probability of changing rating pools of 100%, 48%, or 26%, the rater profile of 39%, 49%, or 50%, the average rating pool size of 5, 10, or 15 officers, and the sorting function of Q'_{i1}, Q'_{i2}, or Q'_{i3}.

The first stage of the NSGS procedure consisted of 25 replications for each of the 81 configurations using the Simio simulation software package. After the first sampling stage, 73 configurations were eliminated from contention as the best configuration. Table 4.2 shows the input and mean responses of the eight configurations remaining for second stage consideration when ω_{mp}, ω_{sp}, and ω_{cp} are -1, -2, and -3, respectively. Based on the results in Table 4.2, increasing the profile constraint generally reduces the number of critical misidentifications. An increase in the number of top evaluations awarded results in the distribution shown in Figure 5.1. An increase in the number of officers receiving zero or five top evaluations, and a corresponding decrease in the number of officers receiving one, two, three, or four top evaluations, reduces the number of critical misidentifications. Table 4.2 also shows that seven of the eight best system configurations had a average rating pool size of 15 officers. The larger rating pools provide a more uniform distribution of performance percentiles that results in fewer misidentifications.

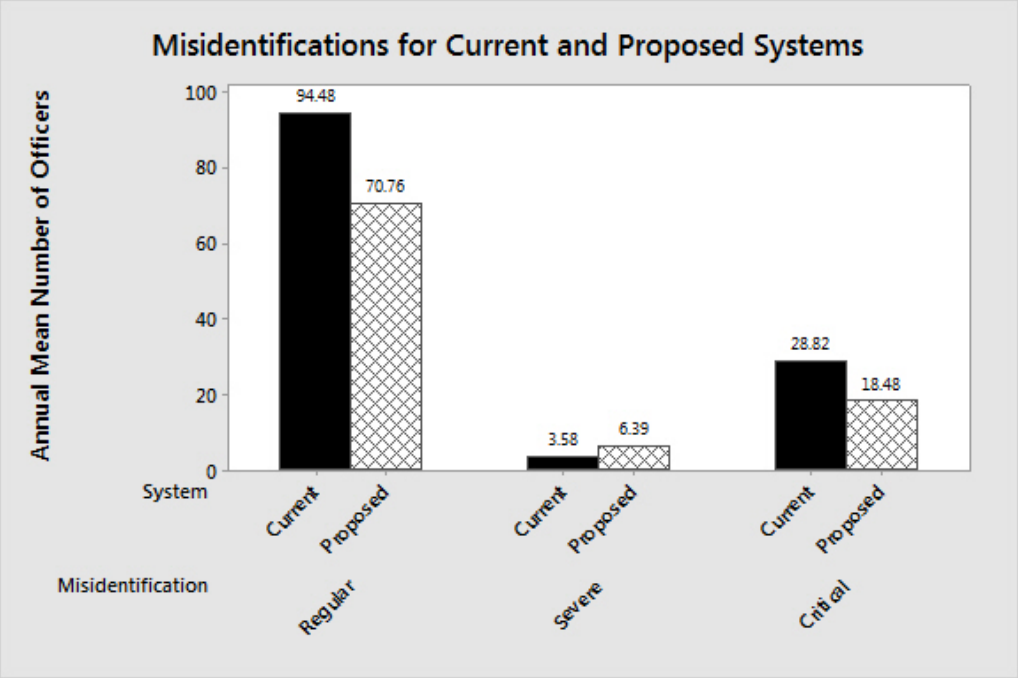


Figure 4.2: A comparison of misidentifications for the current and proposed performance appraisal systems.

Table 4.3 shows the number of replications and the output from the second sampling stage of the NSGS procedure. The highlighted configuration refers to the optimal system with the sorting function Q'_{i2}, officers changing rating pools every year, a rater profile constraint of 0.59, and an average rating pool size of 15. The NSGS procedure took a total of 6,179 replications distributed amongst the 81 system configurations in order to determine the optimal configuration.

Implementation of the KN procedure yielded the same optimal configuration as the NSGS procedure with fewer replications. Table 4.4 shows the total number of replications for each procedure and the output from the optimal configuration. The KN procedure required 60.04% fewer replications than the NSGS procedure. This improved efficiency of the fully-sequential ranking and selection procedure is consistent with the results found by Kim and Nelson (2001).

Table 4.2: Results from the first sampling stage of the NSGS procedure (A: sorting function; B: annual probability of changing rating pools; C: profile constraint; D: average rating pool size; E: misidentifications; F: severe misidentifications; G: critical misidentifications).

Input				Output		
A	B	C	D	E	F	G
Q'_{i2}	1	0.39	15	73.80	5.60	20.44
		0.49	15	74.40	5.92	19.28
		0.59	15	70.04	6.20	17.88
		0.59	10	85.16	8.60	22.04
	0.48	0.39	15	76.80	6.84	23.04
		0.49	15	75.16	8.28	21.36
		0.59	15	73.68	8.44	20.24
	0.26	0.59	15	63.92	12.16	21.08

Table 4.3: Results from the second sampling stage of the NSGS procedure (A: sorting function; B: annual probability of changing rating pools; C: profile constraint; D: average rating pool size).

Input				Replications	Output
A	B	C	D	N_b	$\overline{M}_\ell^{(2)}$
Q'_{i2}	1	0.39	15	555	-149.68
		0.49	15	979	-148.44
		0.59	**15**	**423**	**-139.35**
		0.59	10	523	-166.08
	0.48	0.39	15	588	-156.88
		0.49	15	309	-155.09
		0.59	15	354	-148.25
	0.26	0.59	15	623	-152.37

4.6 Robustness of Responses

The robustness of the results can be determined by calculating the mean response for each configuration under varying values of ω_{mp}, ω_{sp}, and ω_{cp}. The aforementioned ranking and selection procedures involved ω_{mp}, ω_{sp}, and ω_{cp} values of -1, -2, and -3, respectively. That is, severe misidentifications are penalized twice as much as regular misidentifications, while critical misidentifications are

Table 4.4: Comparison of optimal solution found by NSGS and KN procedures (E: misidentifications; F: severe misidentifications; G: critical misidentifications; H: mean response value ($\overline{M}_\ell^{(2)}$ for NSGS, $\overline{M}_\ell(b)$ for KN)).

Method	Total Replications	Output E	F	G	H
NSGS	6179	71.33	6.17	18.55	-139.35
KN	2469	70.76	6.39	18.48	-138.96

penalized three times as much as regular misidentifications. The logical, linear relationship between ω_{mp}, ω_{sp}, and ω_{cp} is in accordance with $\omega_{cp} < \omega_{sp} < \omega_{mp}$. In order to assess the robustness of the results, we tested multiple values of the penalties associated with each form of misidentification.

Each penalty setting PS_c, $c = 1, 2, \ldots, 6$, is detailed in Table 4.5. The penalty for regular misidentifications is fixed at -1 for each of the six penalty settings tested. This provides a baseline for comparison when determining the penalties for severe and critical misidentifications. The first three penalty settings (PS_1, PS_2, and PS_3) maintained a linear relationship between ω_{mp}, ω_{sp}, and ω_{cp}, but had different magnitudes of the penalties associated with each type of misidentification. For PS_4 and PS_5, a non-linear relationship was considered between ω_{mp}, ω_{sp}, and ω_{cp}. Finally, PS_6 penalized each type of misidentification equally.

Table 4.5: Penalty settings (PS) for ω_{mp}, ω_{sp}, and ω_{cp} used in sensitivity analysis.

Penalties	PS_1	PS_2	PS_3	PS_4	PS_5	PS_6
ω_{mp}	-1	-1	-1	-1	-1	-1
ω_{sp}	-2	-3	-5.5	-2	-9	-1
ω_{cp}	-3	-5	-10	-10	-10	-1

The optimal configuration shown in Table 4.3 remains unchanged for each of the six penalty settings. Table 4.6 presents the performance of the five best configurations under the NSGS procedure for varying penalty values. In addition to

the optimal configuration remaining unchanged, all five of the best configurations under the NSGS procedure performed well for varying penalty values.

Table 4.6: Configuration rankings under various penalty settings (A: sorting function; B: annual probability of changing rating pools; C: profile constraint; D: average rating pool size).

Input				Ranking for Penalty Setting (PS)					
A	B	C	D	1	2	3	4	5	6
Q'_{i2}	1	0.39	15	4	3	3	3	4	5
		0.49	15	3	2	2	2	2	4
		0.59	15	1	1	1	1	1	1
	0.48	0.59	15	2	4	5	4	6	3
	0.26	0.59	15	5	7	8	7	9	2

CHAPTER 5

DISCUSSION

5.1 Results

Forced distribution performance appraisal systems produce varied levels of accuracy that are a result of system dynamics, system structure, and rater behavior within the system. The optimal performance appraisal system configuration via the NSGS and KN procedures contains notable deviations from the current configuration. We describe each of the decision variables listed in Constraints (4.2.3)-(4.2.6) and their relationship to current system configuration.

The optimal configuration from Table 4.6 has the same average rating pool size as the current system (15 officers), while the move frequency, profile constraint, sorting functions are varied. Officers currently spend an average of 16.42 months in each assignment, and the optimal configuration has officers spending 12 months in each assignment, the minimal amount of time allowed by Constraint (4.2.3).

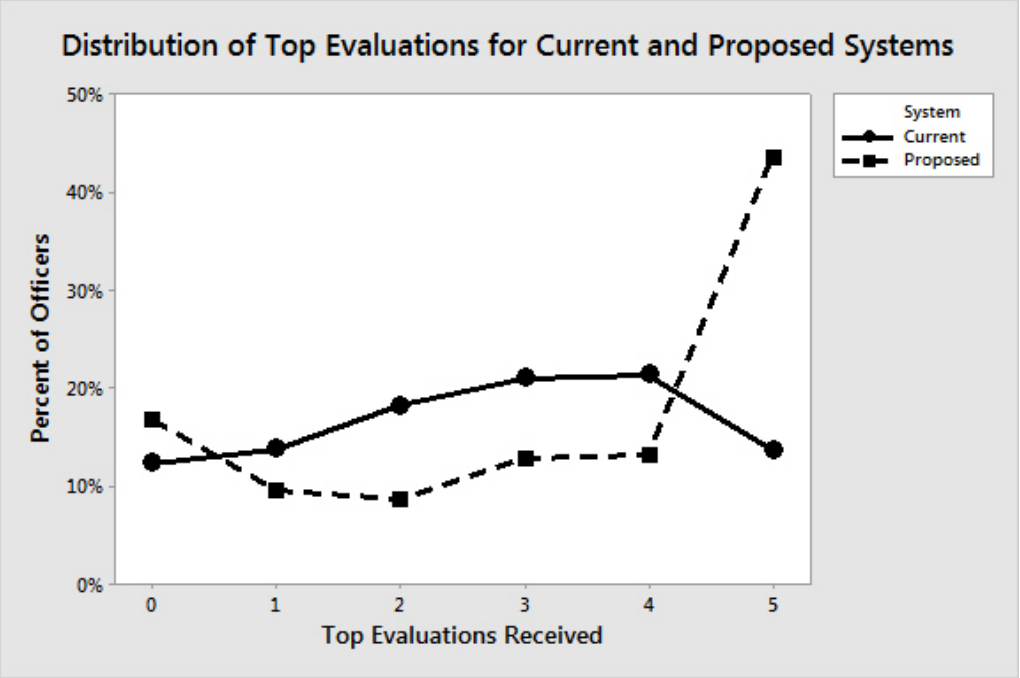


Figure 5.1: The distribution of top evaluations for the current and proposed performance appraisal systems.

The profile constraint of the optimal configuration is 59% as opposed to the current 49%, allowing raters to give more top evaluations than currently allowed by regulation. Finally, the sorting function for the optimal configuration, Q'_{i2}, places less weight on seniority than prevalent of rater behavior in the current system.

Each combination of performance appraisal system variables produces a unique distribution of top evaluations received by the officers in the system. Figure 5.1 shows the distribution of top evaluations for the proposed system compared to the evaluation distribution of the current system. With the profile constraint changed from 49% to 59% in the optimal configuration, an optimal combination of the decision variables produced a distribution that is substantially different than that of the current system. Most notably, over 43% of the officers in the proposed system receive five top evaluations, compared to 14% in the current system.

Just as the distribution of top evaluations in the proposed system differs from that of the current system, the magnitude of each type of misidentification also varies as the decision variables change. Figure 4.2 shows the annual number of misidentifications, severe misidentifications, and critical misidentifications for the current and proposed performance appraisal systems. The proposed system reduces the number of annual regular and critical misidentifications by over 23 and 10 officers, respectively, but slightly increases the annual number of severe misidentifications by approximately three officers. However, since the magnitude of the decrease in critical misidentifications outweighed the magnitude of the increase in severe misidentifications, the proposed configuration demonstrated an improved performance across all penalty settings tested in Table 4.5.

Any change to the performance appraisal system to meet the optimal configuration would likely face various levels of difficulty. Changes in the profile constraint from 49% to 59% can be easily made through regulation. Regulating the average rating pool size and the mean time spent in an assignment is more challenging since not all units are homogeneous. That is, an operational unit will have a different structure than a unit such as the Human Resources Command, and the rating structure may not logically conform to the recommended 15 officers in each rating pool. Moreover, decreasing the amount of time in each assignment would have budget impacts as a result of increasing the frequency of moving officers between duty locations. Rater behavior in the system is likely to pose a challenge to policy makers because imposing restrictions on the number of top evaluations (raters can give in relation to their subordinates' seniority) would be difficult to

implement. Evaluating subsets of officers based on their seniority would introduce additional errors due to a decrease in rating pool sizes. Further, this would also ignore the prevailing assertion that an officer's performance increases as a function of seniority. However, over 70% of the officers in the current system receive a top evaluation in their final year as a major, resulting in an increased number of officers receiving two or three total top evaluations, as shown in Figure 5.1, and subsequently inflates the number of critical misidentifications, as shown in Figure 4.2.

5.2 Future Research

An uneven performance distribution has an effect on the number and severity of misidentifications within a performance appraisal system. To this point, the results are based on a random performance distribution between rating pools. That is, the assignment of an officer to a rating pool is made irrespective of the officer's performance level. However, there are units in the U.S. Army that have the ability to accept or reject officers prior to assignment. These organizations, commonly referred to as *nominative* units, receive a larger number of high-performing officers than other units. Army Regulation 623-3, which dictates evaluation policy, does not make concessions regarding the profile constraints of raters within nominative units. As a result, a disproportionate number of high-performing individuals within the same rating pool increases the number of misidentifications. Figure 5.2 shows the effect of 10% of the rating pools accepting only officers whose

performance is in the top 50% of their peers.

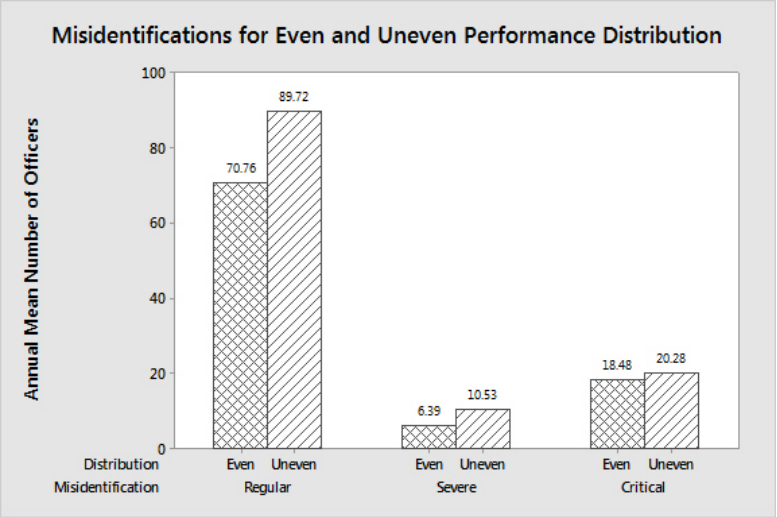


Figure 5.2: A comparison of misidentifications for the proposed performance appraisal system with an even and uneven performance distribution.

The uneven distribution labeled in Figure 5.2 represents the output from a system with the same parameters as the proposed system in Figure 4.2, with the exception that 10% of the rating pools are nominative organizations who only accept officers in the top 50% of their peers. The uneven performance distribution results in a 26.79% increase in regular misidentifications, a 64.79% increase in severe misidentifications, and a 9.74% increase in critical misidentifications. Further research is required to estimate the performance distribution across rating pools, but the initial estimation displayed in Figure 5.2 shows that an uneven distribution of performance levels across units leads to a significant increase in misidentifications.

Further research is required to accurately model the distribution of rating pool sizes. Figure 3.1 shows that approximately 10% of the rating pools for majors facing promotion boards in 2015 and 2016 consisted of 40 or more officers. After

Army Regulation 623-3 directly addressed this practice of pooling, the number of large rating pools should decrease over time. Capturing the effect of policy changes is necessary to validate simulation models for future analysis.

We note that the stated results and conclusions are estimates based on functional area officers, a subset of Army officers, who faced promotion boards in 2015 and 2016. To fully understand the complexities of the U.S. Army performance appraisal system, future research would need to examine the effect and timing of key developmental assignments. Furthermore, the most recent change to the officer evaluation report allows for both a rater and senior rater box check and implements a profile constraint of 49% for both raters and senior raters. As these officers face promotion boards, further research is required to determine the effect of both raters' assessments on both promotion board results and the accuracy of the performance appraisal system.

CHAPTER 6

CONCLUSIONS

The accuracy of an organization's performance appraisal system enhances or limits the ability to retain and promote the highest performing individuals, a critical component of manpower planning and more directly, talent management. This research has addressed the development and analysis of simulation models that integrate system structure and human behavior in order to estimate the effectiveness of the U.S. Army's performance appraisal system. The simulation models serve as tools for policy analysis to determine the extent to which proposed policy accurately identifies the most qualified employees. The results show broad applicability of simulation optimization in the field of manpower modeling and human resource management.

Manpower modeling attempts to match the supply of personnel with the jobs available for them. Original manpower modeling literature focused on developing closed form solutions for determining recruitment and promotions for multi-tiered organizations based on fixed organizational structures and historical attri-

tion rates. Researchers have applied these concepts to military organizations in order to develop retention incentives, restructure retirement benefits, and analyze the military's ability to respond to future conflicts. Military organizations provide large enough samples to draw meaningful conclusions from appropriately aggregated data. However, aggregation does not necessarily mean loss of fidelity regarding individual characteristics and qualifications.

The proposed methodology is not simply building a discrete event simulation of the performance appraisal process. Rather, this research provides a framework to analyze the effect of organizational structure, dynamics, and rater behavior on the organization's ability to identify and promote the most qualified individuals for future job requirements. The proposed framework allows leadership to estimate unintended consequences of policy change by simulating policy changes and comparing the system output to the current system. Moreover, the framework provides a method to recommend personnel policy by optimizing input parameters and regulatory constraints through simulation-optimization techniques.

When applied to the U.S. Army performance appraisal system, this simulation-optimization approach quantifies the classification errors that individuals have intuitively believed without understanding the conditions that exacerbate the effects. The computational results presented using data obtained from the United States Army Human Resources Command demonstrates the potential to reduce the number of errors within the performance appraisal by over 26% and reduce the number of errors likely affecting promotions by over 35%. Incorporating the

quality employee performance into the decision-making process of manpower modeling is a new thread of increasing importance in the fields of military manpower planning and talent management.

REFERENCES

1. 114th Congress, 1st Session (2015), "HR1735: National Defense Authorization Act for Fiscal Year 2016."

2. April, J., Glover, F., and Kelly, J. (2002), "Portfolio Optimization for Capital Investment Projects", *Proceedings of the 2002 Winter Simulation Conference*, 1546-1554.

3. Armitage, A., Schultz, B., Davis, E., and Lykins, L. (2012), "Performance Management Playbook: Managing Critical Performance Challenges, Actionable Strategies Based on Rigorous Research", Institute for Corporate Productivity Technical Report, Seattle, WA.

4. Axlerod, R. (1997), "*Advancing the Art of Simulation in the Social Sciences*", *Simulating Social Phenomena*, eds. R. Conte, R. Hegselmann and P. Terna, Berlin, Germany: Springer-Verlag, **1**, 21-40.

5. Balci, O. (1998), "*Verification, Validation, and Testing*", *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, ed.

J. Banks, New York, NY: John Wiley and Sons, Inc., 335-393.

6. Banks, J. (1998), "*Principles of Simulation*", *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, ed. J. Banks, New York, NY: John Wiley and Sons, Inc., 3-30.

7. Banner, D.K., and Cooke, R.A. (1984), "Ethical Dilemmas in Performance Appraisal", *Journal of Business Ethics*, **3**(4), 327-333.

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

9. 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*, **45**(4), 305-324.

10. Bechhofer, R., Santner, T., and Goldsman, D. (1995), *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, New York, NY: John Wiley and Sons, Inc.

11. Bellman, R. (1954), "The Theory of Dynamic Programming", *Bulletin of the American Mathematical Society*, 513-515.

12. Better, M., Glover, F., Kochenberger, G., and Wang, H. (2008), "Simulation Optimization: Applications in Risk Management", *International Journal of Information Technology & Decision Making*, **7**(4), 571-587.

13. Bjerke, D.G., Cleveland, J.E., Morrison, 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.

14. Boudreau, J.W. (2004), "Organizational Behavior, Strategy, Performance, and Design in Management Sciences", *Management Science*, **50**(11), 1463-1476.

15. Bres, E.S., Burns, D., Charnes, A., and Cooper, D.D. (1980), "A Goal Programming Model for Planning Officer Accessions", *Management Science*, **26**(8), 773-783.

16. Budiansky, S. (2013), *Blackett's War*, Vintage Books, New York, NY.

17. Carroll, S.J., and Schneier, C.E. (1982), *Performance Appraisal and Review Systems: The Identification, Measurement, and Development of Performance in Organizations*, Glenview, IL: Scott, Foresman, and Company.

18. Carson, Y., and Maria, A. (1997), "Simulation Optimization: Methods and Applications", *Proceedings of the 1997 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 118-126.

19. Chang, J.R., Cheng, C.H., and Chen, L.S. (2007), "A Fuzzy-Based Military Officer Performance Appraisal System", *Applied Soft Computing*, **7**(3), 936-945.

20. Cheese, P. (2010), "Talent Management for a New Era: What We Have Learned from the Recession and What We Need to Focus On Next", *Human Resource Management International Digest*, **18**(3), 3-5.

21. Civil Service Reform Act (1978), Public Law 95-454, S 2640, October 13, 1978.

22. 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*, **37**(1), 5-18.

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

24. Dabkowski, M.F., Huddleston, S.H., Kucik, P., and Lyle, D. (2010), "Shaping Senior Leader Officer Talent: How Personnel Management Decisions and Attrition Impact the Flow of Army Officer Talent Throughout the Officer Career Model", *Proceedings of the 2010 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 1407-1418.

25. Dabkowski, M.F., Huddleston, S.H., Kucik, P., and Lyle, D. (2011), "Shaping Senior Leader 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", *Proceedings of the 2011 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 2466-2477.

26. Dailey, J.T. (1958), "Prediction of First-cruise Reenlistment Rate", *Operations Research*, **6**(5), 686-692.

27. David, H.A. (1981), *Order Statistics*, New York, NY: John Wiley and Sons, Inc.

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

29. Department of the Army (2014), *Department of the Army Pamphlet 600-3: Commissioned Officer Professional Development and Career Management*, December, 2014.

30. Department of the Army (2015), *Army Regulation 623-3: Evaluation Reporting System*, November, 2015.

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

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

33. Deputy Assistant Secretary of the Army - Budget (2017), "Fiscal Year (FY) 2018 President's Budget Submission", *Army Military Personnel; Justification Book*, 124-129, accessed August 30, 2017 from https://www.asafm.army.mil/documents/BudgetMaterial/fy2018/mpa.pdf.

34. Elmuti, D., Kathawala, Y., and Wayland, R. (1992), "Traditional Performance Appraisal Systems: The Deming Challenge", *Management Decision*, **30**(8), 42-48.

35. Feldman, J.M. (1986), "Instrumentation and Training for Performance Appraisal: A Perceptual Cognitive Viewpoint", *Research in Personnel and Human Resources Management*, **4**, 45-99.

36. Fisher, F.M. and Morton, A.S. (1967), "The Costs and Effectiveness of Reenlistment Incentives in the Navy", *Operations Research*, **15**(3), 373-387.

37. Fu, M.C. (2001), "Simulation Optimization", *Proceedings of the 2001 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 53-61.

38. Fu, M.C., Glover, F.W., and April, J. (2005), "Simulation Optimization: A Review, New Developments, and Applications", *Proceedings of the 2005 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 83-95.

39. Gass, S.I., and Assad, A.A. (2005), *An Annotated Timeline of Operations Research*, New York, NY: Spencer Science and Business Media, Inc.

40. Georgiou, A.C., and Tsantas, N. (2002), "Modelling Recruitment Training in Mathematical Human Resource Planning", *Applied Stochastic Models in Business and Industry*, **1**, 53-74.

41. Goldsman, D., and Nelson, B.L. (1994), "Ranking, Selection and Multiple

Comparisons in Computer Simulation", *Proceedings of the 1994 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 192-199.

42. Goodeve, C. (1948), "Operational Research", *Nature*, **161**, 377-384.

43. Groves, K.S. (2011), "Talent Management Best Practices: How Exemplary Health Care Organizations Create Value in a Down Economy", *Health Care Management Review*, **36**(3), 227-240.

44. Hall, A.O. (2009), "Simulating and Optimizing: Military Manpower Modeling and Mountain Range Options", Ph.D. Dissertation, University of Maryland, College Park, MD.

45. Hangos, K.M., and Cameron, I.T. (2001), *Process Modelling and Model Analysis*, Academic Press, London, United Kingdom.

46. 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*, **61**, 768-779.

47. Henry, T.M., and Ravindran, A.R. (2005), "A Goal Programming Application for Army Officer Accession Planning", *INFOR: Information Systems and Operational Research*, **43**(2), 111-119.

48. Holt, J. (2009), *A Pragmatic Guide to Business Process Modelling*, London, United Kingdom: BCS Learning and Development Ltd.

49. Howe, M.J., Davidson, J.W., and Sloboda, J.A. (1998), "Innate Talents: Reality of Myth", *Behavioural and Brain Sciences*, **21**(3), 399-442.

50. Ikonen, E., and Najim, K. (2002), *Advanced Process Identification and Control*, New York, NY: Marcel Dekker, Inc.

51. Imperato, G. (1998), "Tales of Tomorrow", *Fast Company*, **17**, 147.

52. Kahneman, D. (2000), *Thinking Fast and Slow, 1st Edition*, New York, NY: Farrar, Straus, and Giroux.

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

54. Kim, S.-H., and Nelson, B.L. (2001), "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation", *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **11**(3), 251-273.

55. Kinstler, D.P., Johnson, R.W., Richter, A., and Kocher, K. (2008), "Navy Nurse Corps Manpower Management Model", *Journal of Health Organization and Management*, **22**(6), 614-626.

56. Kleijnen, J.P.C., and Sargent, R.G. (2000), "A Methodology for Fitting and Validating Metamodels in Simulation", *European Journal of Operational Research*, **120**, 14-29.

57. Kozlowski, S.W.J., Chao, G.T., and Morrison, R.F. (1998), "*Games Raters Play: Politics, Strategies, and Impression Management in Performance Appraisal.*", *Performance Appraisal: State of the Art in Practice*, ed. J.W. Smither, San Francisco, CA: Jossey-Bass Publishers, 163-205.

58. Kwinn Jr., M.J., and Phelan Jr., R.G. (1996), "Management of Personnel Policies to Increase the Stability of Patriot Crew Members and Their Families: A Simulation Approach", *Proceedings of the 1996 Winter Simulation Conference*, Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., 926-933.

59. Lauer, M.S. (2012), "From Hot Hands to Declining Effects: The Risks of Small Numbers", *Journal of the American College of Cardiology*, **60**, 72-74.

60. Law, A.M. (2015), *Simulation Modeling & Analysis*, 5th ed., New York, NY: McGraw-Hill Education.

61. Lesinski, G., Pinter, J., Kucik, P., and Lamm, G. (2011), "Officer Accessions Flow Model", Technical Report (DSE-TR-1103), Operations Research Center of Excellence, U.S. Military Academy, West Point, NY.

62. Levitin, A. (2003), *Introduction to the Design and Analysis of Algorithms*, Boston, MA: Addison-Wesley.

63. Lew, A., and Mauch, H. (2007), *Dynamic Programming: A Computational Tool*, Berlin, Germany: Springer.

64. Lopez, C.T. (2011), "Box Check Returns to Company-grade OERs', *Army New Service*, accessed January 28, 2018 from https://www.army.mil/article/65755/.

65. McDonnell, A., Collings, D.G., Mellehi, K., and Schuler, R. (2017), "Talent Management: A Systematic Review and Future Prospects", *European*

Journal of International Management, **11**(1), 86-128.

66. McGinnis, M.L., and Fernandez-Gaucherand, E. (1994), "Resource Scheduling for the United States Army's Basic Combat Training Program", *IEEE International Conference on Systems, Man, and Cybernetics*, **1**, 553-558.

67. McGinnis, M.L., Kays, J.L., and Slaten, P. (1994), "Computer Simulation of U.S. Army Officer Professional Development", *Proceedings of the 1994 Winter Simulation Conference*, 813-820.

68. McGregor, D. (1957), "An Uneasy Look at Performance Appraisals", *Harvard Business Review*, **35**(3), 89-94.

69. Meyer, H.H. (1980), "Self-Appraisal of Job Performance", *Personnel Psychology*, **33**, 291-296.

70. Mohrman Jr., A.M., Resnick-West, S.M., and Lawler, E.E. (1989), *Designing Performance Appraisal Systems: Aligning Appraisals and Organizational Realities*, San Francisco, CA: Jossey-Bass Publishers.

71. Morse, P.M., and Kimball, G.E. (1951), *Methods of Operations Research*, New York, NY: John Wiley and Sons, Inc.

72. Murphy, K.R., and Cleveland, J.N. (1995), *Understanding Performance Appraisal: Social, Organizational, and Goal-based Perspectives*, Thousand Oaks, CA: SAGE Publications, Inc.

73. Nelson, B.L., Swann, J., Goldsman, D., and Song, W. (2001), "Simple Procedures for Selecting the Best Simulated System when the Number of Alter-

natives is Large", *Operations Research*, **49**(6), 950-963.

74. Office of the Under Secretary of Defense for Personnel and Readiness (2016), "Department of Defense Instruction 1400.25: DoD Civilian Personnel Management System: Performance Management and Appraisal System", Vol. 431, 11.

75. Office of the Under Secretary of Defense, Chief Financial Officer (2015), "United States Department of Defense Fiscal Year 2016 Budget Request Overview", accessed February 15, 2016 from www.comptroller.gov.

76. Odierno, R.T. (2015), "Leader Development and Talent Management: The Army Competitive Advantage", *Military Review*, July-August 2015, 98-108.

77. Ozdemir, O. (2013), "A Two-phase Multi Criteria Dynamic Programming Approach for Personnel Selection Process", *Problems and Perspectives in Management*, **12**(2), 98-108.

78. Peck, C.A. (1984), *Pay and Performance: The Interaction of Compensation and Performance Appraisal*, Research Bulletin no. 155, Conference Board, New York, NY.

79. Price, W.L., and Piskor, W.G. (1972), "The Application of Goal Programming to Manpower Planning", *INFOR: Information Systems and Operational Research*, **10**(3), 221-231.

80. Rao, P.P. (1990), "A Dynamic Programming Approach to Determine Optimal Manpower Recruitment Policies", *Journal of the Operational Research*

Society, **41**(10), 365-381.

81. Rinott, Y. (1978), "On Two-stage Selection Procedures and Related Probability-inequalities", *Communications in Statistics*, A7, 799-811.

82. Rostker, B., Thie, H. Lacy, J., Kawata, J., and Purnell, S. (1993), "The Defense Officer Personnel Management Act of 1980: A Retrospective Assessment", Technical Report (R-4246-FMP), RAND, Santa Monica, CA.

83. Shutler, M., and Storbeck, J. (2002), "Performance Management", *The Journal of the Operational Research Society*, **53**(3), 245-246.

84. Sisson, E.D. (1948), "Forced Choice: The New Army Rating", *Personnel Psychology*, **1**, 365-381.

85. Smith, P.C., and Goddard, M. (2002), "Performance Management and Operational Research: A Marriage Made in Heaven?", *The Journal of the Operational Research Society*, **53**(3), 247-255.

86. Sparrow, P.R., and Makram, H. (2015), "What is the Value of Talent Management? Building Value-driven Processes within a Talent Management Architecture", *Human Resource Management Review*, **25**(3), 249-263.

87. Staugas, l., and McQuitty, L.L. (1950), "A New Application of Forced-choice Rating", *Personnel Psychology*, **3**(4), 413-424.

88. Stevenson, C., and DiRomualdo, T. (2013), "Performance Management 2013: Still Waiting for Real Change", Institute for Corporate Productivity Technical Report, Seattle, WA.

89. Swailes, S. (2013), "The Ethics of Talent Management", *Business Ethics: A European Review*, **22**(1), 32-46.

90. Troitzsch, K.G. (1997), "Social Science Simulation - Origins, Prospects, Purposes", *Simulating Social Phenomena*, eds. R. Conte, R. Hegselmann and P. Terna, Springer-Verlag, **1**, 41-54.

91. Ulrich, D., and Smallwood, N. (2012), "What is Talent?", *Leader to Leader*, **63**, 55-61.

92. Wang, J. (2005), "A Review of Operations Research Applications in Workforce Planing and Potential Modelling of Military Training", Australian Department of Defence Science and Technology Organisation Technical Report, Edinburgh, Australia.

93. Wardynski, C., Lyle, D.S., and Colarusso, M.J. (2009), "Talent: Implications for a U.S. Army Officer Corps Strategy", *Strategic Studies Institute: Officer Corps Strategy Monograph Series*, Vol. 2.

94. Wardynski, C., Lyle, D.S., and Colarusso, M.J. (2010), "Towards a U.S. Army Officer Corps Strategy for Success: Retaining Talent", *Strategic Studies Institute: Officer Corps Strategy Monograph Series*, Vol. 3.

95. Wessels, J., and van Nunen, J.A.E.E. (1976), "FORMASY FOrecasting and Recruitment in MAnpower Systems", *Statistica Neerlandica*, **30**(4), 173-193.

96. Wiese, D.S., and Buckley, M.R. (1998), "The Evolution of the Performance Appraisal Process", *Journal of Management History*, **4**(3), 233-249.

97. WorldatWork and Sibson Consulting (2010), "2010 Study on the Current State of Performance Management", Technical Report, Washington, DC.

98. Zais, M.M. (2014), "Simulation-Optimization, Markov Chain and Graph Coloring Approaches to Military Manpower Modeling and Deployment Sourcing", Ph.D. Dissertation, University of Colorado, Boulder, CO.

99. Zanakis, S.H., and Maret, M.W. (1980), "A Markov Chain Application to Manpower Supply Planning", *Journal of the Operational Research Society*, **31**(12), 1095-1102.

100. Zhu, Y. (2001), *Multivariable System Identification for Process Control*, Kidlington, United Kingdom: Elsevier Science Ltd.

CURRICULUM VITAE

NAME: Lee A. Evans

ADDRESS: Department of Industrial Engineering
JB Speed School of Engineering
University of Louisville
132 Eastern Parkway
Louisville, KY 40292

DOB: San Antonio, TX - December 30, 1977

EDUCATION
& TRAINING: B.S., Engineering Management
United States Military Academy
West Point, NY
1996-2000

M.S., Operations Research
Georgia Institute of Technology
Atlanta, GA
2007-2009

Ph.D., Industrial Engineering
University of Louisville
Louisville, KY
2015-2018

PROFESSIONAL
SOCIETIES: INFORMS
MORS
Alpha Pi Mu
Pi Mu Epsilon

PUBLICATIONS: Evans, L.A., and Bae, K.-H.G. (Submitted 2018), "U.S. Army Performance Appraisal Policy Analysis: A Simulation Optimization Approach," *Journal of Defense Modeling and Simulation.*

Evans, L.A., and Bae, K.-H.G. (2018), "Simulation-Based Analysis of a Forced Distribution Performance Appraisal System," *Journal of Defense Analytics and Logistics.*

Evans, L.A., Bae, K.-H.G., and Roy, A. (2017), "Single and Multi-Objective Parameter Estimation of a Military Personnel System," *Proceedings of the 2017 Winter Simulation Conference*,

4058-4069.

Bae, K.-H.G., Evans, L.A., and Summers, A. (2016), "Lean Design and Analysis of a Milk-Run Delivery System: Case Study," *Proceedings of the 2016 Winter Simulation Conference*, 2855-2866.

Evans, L.A., Bodenheim, E.H., and Fawson, L. (2013), "The Officer Assignment Process: From Science to Art," *1775: The Journal of the Adjutant General Regimental Association*, Winter 2013-2014, 50-54.

Evans, L.A., and Weld, C.E. (2011), "Assessing and Improving Students' Fundamental Mathematical Skills," *Mathematica Militaris*, Volume 20, Issue 2, 2-9.