Evaluating Our Evaluations
Recognizing and Countering Performance Evaluation Pitfalls

Lt. Col. Lee A. Evans, PhD, U.S. Army
Lt. Col. G. Lee Robinson, PhD, U.S. Army

Selecting the right person for the right job at the right time is a persistent challenge faced by organizations. Performance evaluations are a fundamental component of selection processes, and their use in the Army is nearly as old as the service itself. Some early evaluation systems consisted of a list of officers in a regiment with observations noted for each ranging from “a good-natured man” to “merely good—nothing promising” to “a man of whom all unite in speaking ill.”

While our current evaluation form adds a bit more science to the art of performance evaluation, a constant in the Army’s performance evaluation system is the reliance on raters to render their judgment on the potential of a subordinate for service at higher levels.

Raters need to be better equipped to exercise these judgments. While we recognize the calls for personnel management reform and the initiatives underway to better manage the Army’s talent, our purpose is not to add another voice to these suggestions for structural changes to the Army’s evaluation system. Instead, we focus on the process of discretionary judgment exercised by raters that is and will continue to be an integral part of performance evaluation. Our aim is to recognize the structural and cognitive biases inherent in our evaluation system and provide recommendations to help senior raters more objectively evaluate their subordinates.

While we think the importance of this topic is self-evident, educating raters on the potential for bias in their evaluations is especially important in the type of rating system used by the Army. This system places great emphasis on the person serving as the senior rater. Although the evaluation forms include assessments from raters and sometimes intermediate raters, the senior rater comments are widely acknowledged to carry the most weight for promotion and selection decisions due to the small amount of time available to evaluate a soldier’s file. Most positions involve work that is highly interdependent on other members of the organization, which places a considerable demand on raters to assess and articulate how much an individual contributed to the output of the group.

While the performance of an officer is undoubtedly important to his or her chances for promotion or selection, the abilities of the officer’s senior rater to convey the level of this performance through an evaluation is also vital to talent management. Previous studies demonstrate that exposure to a high-quality mentor increases
an officer's likelihood of an early promotion to major by 29 percent, perhaps because high-quality mentors are skilled at communicating their protégé's potential in their performance evaluations. Equipping raters to make their best possible judgments of subordinates and clearly articulating these judgments is vital to fostering a meritocratic Army talent management system.

**Evaluating the Performance Evaluation Tool: Structural Biases in the Department of the Army Form 67**

In 1922, the Army introduced a formalized performance appraisal system, the War Department Adjutant Generals Office (WD AGO) Form 711, *Efficiency Report*, rebranded two years later as the WD AGO Form 67, to assess officers in the domains of physical qualities, intelligence, leadership, personal qualities, and general value to the service. Since 1922, the Army modified DA Form 67 ten times; the most recent iteration was the DA Form 67-10 series (hereafter referred to collectively as DA Form 67-10). Each iteration of the officer evaluation form contained nuanced approaches to segment the population in order to accurately represent the spectrum of officer performances from the highest performing officers to those who should not be retained in the service. DA Form 67-10 uses a forced distribution technique where senior raters of lieutenant colonels and below can award “most qualified” evaluations to fewer than half of their subordinates. (For comparison, an example of the 1934 efficiency report format is shown on pages 94–95 to highlight the perennial challenges the Army has faced over time in capturing and expressing an effective and fair means of comparing the performances of officers.) Forced distribution rating systems have been common in the Department of Defense and the civilian sector because of the problem of appraisal distortion in the absence of forced distribution. For example, prior to implementing a forced distribution performance appraisal system, the U.S. Navy saw the majority of its officers rated in the top 1 percent. In theory, forced distribution decreases ratings inflation and provides the means for a variety of human resources decisions, including promotion, training, and assignment of personnel.

However, even under a best-case scenario (with the absence of cognitive biases), system structure induces error in a forced distribution performance appraisal system. Allan Mohrman alluded to this problem in his argument that forced distribution systems should be applied to large enough groups of employees, specifically over fifty. While he failed to provide mathematical support for this number, his argument relies on the statistical qualities of large sample sizes. For example, if a reasonably large sample, typically \( n > 30 \), is drawn from a population with a normal distribution, the sample mean and the standard deviation of the sample are nearly indistinguishable from that of the population. In the context of officer performance and potential, assuming both are normally distributed, this suggests that larger samples of officers will provide a more accurate representation of performance levels across the force. While larger samples are typically a good representation of performance level distribution, they are in direct conflict with the concept of pooling introduced by Army Regulation (AR) 623-3, *Evaluation Reporting System*.

AR 623-3 defines pooling as “elevating the rating chain beyond the senior 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.” The word “pooling” appears more than ten times in the most recent version of AR 623-3, which states that pooling runs counter to the intent of the evaluation system and erodes soldiers’ confidence in the fairness and impartiality of their leaders.

Creating a rating scheme that minimizes the number of subordinates under each rater ideally allows raters to have an intimate knowledge of the strengths and weaknesses of the soldiers they rate. The idea of an organizational structure that limits the number of subordinates under a rater’s span of control is also a common practice in the civilian sector. The manager-to-employee ratio across industries worldwide is approximately 1:4 for companies with five hundred or fewer employees and 1:9 for companies with greater than five hundred employees. While there are many sound reasons that the Army seeks to decrease a rater’s span of control, an often overlooked downside of this practice is the presence of errors resulting from a forced distribution system, especially in small rating pools. According to AR 623-3, a senior rater should award “most qualified” evaluations to the top one-third of officers, and the number of “most qualified” evaluations they award must be less than 50 percent of the total number of evaluations he or she writes.

With a few simplifying assumptions, such as officers distributed randomly into rating pools of five
and the raters having perfect clarity on whether a subordinate is a top one-third officer, the hypergeometric distribution (as explained below) provides insight into the mathematical pitfalls of a forced distribution performance appraisal system.

The hypergeometric distribution has three parameters: \( N \), \( R \), and \( n \). The parameter \( N \) represents the number of items in the population, \( R \) represents the number of “successes,” and \( n \) is the sample size drawn from the population. Using this nomenclature, we can determine that the random variable is \( X \sim \text{Hypergeometric}(N, R, n) \) and calculate the probability that \( X \) (in our case, the number of “most qualified” officers in a rating pool) takes on particular, discrete values.

For example, if there are five thousand officers of a particular rank, 1,667 of them would be considered the top one-third based on established criteria. We can calculate the probability of receiving exactly \( x \) top one-third officers in a group of \( n \) size. If we assume a pool size of five officers, we would use \( X \sim \text{Hypergeometric}(5000, 1667, 5) \) to calculate the probability that we receive exactly \( x \) top one-third officers in our rating pool, notationally \( P(X = x) \). That is, \( P(X = 2) \) represents the probability that exactly two top one-third officers were assigned to a rating pool of five. In fact, \( P(X = 2) = 0.329 \), meaning there is a 32.9 percent chance that there would be exactly two top one-third officers in a rating pool of five, assuming officers are randomly distributed into ratings pools. Thus, given the current profile constraint of less than 50 percent, raters could only award two “most qualified” evaluations to a pool of five officers.

The rater’s ability to discern the two top one-third performers is affected by cognitive biases, but mathematically, the rater may be obligated to award an evaluation that is not commensurate with a subordinate’s level of performance due to forced distribution requirements. For example, if a rater has a pool size of five, but has more than two top one-third performers, at least one rated officer will receive an inaccurate evaluation due to the rater’s profile constraint. We can calculate this expected annual error with \( E[\text{Annual Error}] \). Notationally, for a rating pool of five officers, this is represented by \( E[\text{Annual Error}] = \sum_{i=1}^{5} (i - 2) P(X = i) = P(X = 3) + 2P(X = 4) + 3P(X = 5) \). That is, when there are three top one-third officers in a rating pool of five, one officer is adversely affected by the

Creating a rating scheme that minimizes the number of subordinates under each rater ideally allows raters to have an intimate knowledge of the strengths and weaknesses of the soldiers they rate.

Addressing Structural Biases

We suggest three ways to counter structural biases. First, senior raters should follow the guidance in AR 623-3 and reserve “most qualified” evaluations for the top one-third officers. This requires a discerning eye, and as previously mentioned, will result in an expected annual error of about one officer per rating pool every four years for a rating pool of five officers. According to the U.S. Army Human Resources Command, “the limitation of less than 50% translates to an average use of 37–42% depending on the grade (of the rated officer).” Within this relatively small range, there is a significant difference in the expected annual error.

If a senior rater uses the top 37 percent of officers as the cutoff for most “qualified” evaluations, it would result in an expected annual error of 0.340 whereas a 42
percent threshold increases the expected annual error to 0.469. As seen in figure 1, higher thresholds for what percentage of officers should receive a “most qualified” evaluation result in monotonically higher than expected annual errors. However, senior raters who place these thresholds below those of other raters disadvantage some of their subordinates who would have received “most qualified” evaluations in other rating pools. Therefore, a senior rater would want to award a similar percentage of “most qualified” evaluations as other senior raters across the Army to ensure his or her subordinates are not disadvantaged but low enough to prevent instances where the number of “most qualified” officers within their rating pools exceeds the profile constraint.

Second, we recommend senior raters have a multiyear focus and refrain from maximizing the number of “most qualified” evaluations awarded each year. The U.S. Human Resources Command stated that the 37–42 percent use of “most qualified” evaluations by senior raters is “indicative of senior raters correctly retaining a buffer.”16 This guidance assumes that anything less than 50 percent constitutes a buffer. However, figure 2 (on page 93) shows that the maximum allowable percentage of “most qualified” evaluations does not remain above 42 percent until a senior rater completes twenty-five evaluations. For example, if a senior rater completes eight evaluations, at most, three of them can be “most qualified” evaluations, putting the senior rater profile usage at 37.5 percent. If the senior rater kept a buffer of just one evaluation, the profile usage drops to 25 percent.

Maximizing the number of “most qualified” evaluations awarded often results in either a Type I or Type II error. In the context of performance appraisals, a Type I error is incorrectly identifying an officer as most qualified, whereas Type II error is not identifying a most qualified officer as such. If a senior rater has a rating pool of five officers and is predetermined to award the maximum of two top evaluations, there is only a 34.6 percent chance that there are exactly two top 40 percent officers in a pool of randomly distributed officers. There is a 33.7 percent chance that there are fewer than two top 40 percent officers, leading to a Type I error, and a 31.7 percent chance there are more than two top 40 percent officers, leading to a Type II error. A senior rater’s profile constraint can

---

**Figure 1. Expected Annual Error as a Function of a Senior Rater’s “Most Qualified” Threshold**

![Graph showing expected annual error as a function of a senior rater's most-qualified threshold.](image-url)
induce a Type II error, but a Type I error is caused by either cognitive biases or conscious decisions.

A conscious decision to award a “most qualified” evaluation to an undeserving officer can have compounding effects since rating profiles are cumulative. We analyze this effect by calculating the expected two-year error. If a senior rater plans to maximize the number of “most qualified” evaluations awarded, presumably off of a top 40 percent standard, it will result in an expected annual error of 0.415 and an expected two-year error of 0.830 for a pool size of five. However, if a senior rater can use the top one-third standard for awarding “most qualified” evaluations, there will be an expected annual error of 0.259 and an expected two-year error of 0.416.

The reason that the expected two-year error is not double that of the expected annual error is that if there is only one top one-third officer in the rating pool the first year, the senior rater can award up to three “most qualified” evaluations the second year. Similarly, if there are no top one-third officers in the rating pool the first year, a senior rater can award up to four “most qualified” evaluations the second year. In summary, by resisting the urge to award the maximum allowable number of top evaluations each year and maintaining a top one-third standard, senior raters can reduce Type II errors by nearly 50 percent. Consequently, coaching officers to have a multiyear focus is especially important since recent research shows how an officer’s seniority affects the evaluations they receive in the evaluation process.17

Third, consistent with AR 623-3, we recommend that senior raters structure rating schemes to provide flexibility to reward the best subordinates. When discussing the establishment of rating chains, AR 623-3 provides general guidance, such as commanders rating commanders, and prohibits the practice of pooling. However, it gives organizations the latitude to establish and publish their rating scheme at the beginning of each period. While the recommended size of rating pools cannot be generalized across nonhomogeneous units, organizations should establish rating chains that do not disadvantage officers at each grade level.

For example, increasing our sample rating pool of five officers to ten officers decreases both the expected annual error and the expected annual two-year error. As previously stated, using the criteria of top one-third officers deserving “most qualified” evaluations, the expected annual error for a pool size of five is 0.259 and the expected two-year error is 0.416. Doubling the size
### EFFICIENCY REPORT

**Fissell, John T.**

Bn. Commander, 3rd Bn., 100th Inf.  
**0746642**  
**Lt. Col. 100th Infantry**  

A. OFFICER REPORTED UPON  

B. PERIOD COVERED BY THIS REPORT  

C. STATIONS AT WHICH HE SERVED  

D. CONSIDER CAREFULLY THE DEFINITIONS, KEEP THEM IN MIND WHEN RATING, TAKING INTO CONSIDERATION HIS LENGTH OF SERVICE AND THE OPPORTUNITIES AFFORDED HIM, WHICH MIGHT HAVE A BEARING UPON HIS PERFORMANCE OF DUTY, PERSONAL CHARACTERISTICS, OR PROFESSIONAL QUALIFICATIONS.

<table>
<thead>
<tr>
<th>Duty</th>
<th>Months</th>
<th>Manner of performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bn. Commander, 3rd Bn., 100th Inf. (Prin. duty)</td>
<td>7/20/32</td>
<td>Excellent</td>
</tr>
<tr>
<td>Sick in hospital</td>
<td>7/30</td>
<td></td>
</tr>
</tbody>
</table>

E. DUTY ON WHICH HISTORY OF SERVICE IS BASED. Where possible show duration of each months. Example: Co. Commander, ordinary garrison training, 6 mos. Summary court, 5 mos. Brig. Adj. prepared training schedules, Supply Officer. In describing the manner of performance of duty, use one of six classifications as given under D, and consider carefully the obstacles encountered by the individual in the performance of each duty listed. THE OPINIONS EXPRESSED UNDER "MANNER OF PERFORMANCE" ARE BASED ON—

- Intimate Daily Contact
- Frequent or Infrequent Observation of the Results of His Work
- Academic Ratings

F. Types of reports used in determining grade and value of services rendered during this period.

1. Holding officers and men,
2. Performance of field duties,
3. Administrative and executive duties,
4. As an instructor,
5. Training troops,
6. Tactical handling of troops (units appropriate to officer's grade).

G. Enter on lines below any outstanding special duties or qualifications of value in the military service. MAKE NO ENTRIES EXCEPT WHERE A STATEMENT IS BASED ON PERSONAL OBSERVATION OR OFFICIAL REPORTS DURING PERIOD COVERED BY THIS REPORT. Show pilot and/or observer ratings of Air Corps officers here.

<table>
<thead>
<tr>
<th>Handing officers and men</th>
<th>Statistically</th>
<th>Very satisfactory</th>
<th>Satisfactory</th>
<th>Regular</th>
<th>Excellent</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance of field duties</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrative and executive duties</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>As an instructor</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training troops</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tactical handling of troops (units appropriate to officer's grade)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

H. To what degree has he exhibited the following qualifications? Consider him in comparison with officers of his grade and indicate your estimate by marking X in the appropriate rectangle. (See par. D above.)

1. Physical activity (ability to work rapidly),
2. Physical endurance (capacity for prolonged service),
3. Military bearing and manners (fidelity of character; and good aptitude),
4. Attention to duty (the trait of working thoroughly and conscientiously),
5. Cooperation (acting jointly and effectively with another, military or civilian, to attain a designated objective),
6. Initiative (the trait of seizing needed work or taking appropriate action on his own responsibility in absence of orders),
7. Inteligence (the ability to understand readily new ideas or instructions),
8. Force (the trait of maintaining energy and resolution that which an occasion or a belief demands, right or wrong),
9. Judgment and common sense (the ability to think clearly and arrive at logical conclusions),
10. Leadership (the trait of direct, exact, and influence others in details lines of action or movement and still maintain high work)

W. D., A. G. O. Form No. 67—July 1, 1939.

FIGURE 35.

(From published in Technical Manual 12-250, Administration, 10 February 1942)

---

Sample of U.S. Army Efficiency Report from 1936
Sample of U.S. Army Efficiency Report from 1936 (continued)
of the rating pool to ten officers while maintaining the
top one-third most qualified officer threshold drops
the expected two-year error to 0.364. Since the expect-
ed two-year error is for two years of officers in a pool
size of ten, we can compare it to the expected two-year
of cognitive bias can make a difference in the iden-
tification and selection of officers with the greatest
potential for service at higher levels. Stated differ-
tently, the more bias we can divest from evaluations,
the better positioned selection boards will be to make
error for a pool size of five by dividing by two. Doubling
the rating pool size from five to ten thus results in a 56
percent decrease in Type II errors.

**Evaluating the Evaluator:**
**Cognitive Biases**

As evidenced in the previous section, there are
structural biases introduced by the DA Form 67-10 that
make it difficult for raters to consistently reward the best
officers. In addition to these structural biases, because of
the discretionary nature of performance evaluation, there
are also cognitive biases that may affect the judgment of
senior raters. We focus on five cognitive biases that may
lead to a difference between the performance of an officer
and how this performance translates to the potential
described by a senior rater in an evaluation report.

A cognitive bias occurs when a rater unknowingly
renders judgments that are unrelated to an officer’s
performance. Because raters have great discretion
in how they articulate the potential of an officer in
an evaluation, cognitive biases have the potential to
influence the enthusiasm they use to describe a soldier
in the narrative portion of the report.

These choices are especially important because there is likely a small talent differential between offi-
cers just above and just below the cutline in promo-
tion and selection boards. There is anecdotal evidence
to support this point from officers who served on
promotion boards, but we also see empirical support
for small differences between primary and alternate
selectees in other fields. Since selection boards have
little time to review files and consider a relatively
minimal amount of information, reducing the effects

The more bias we can divest from evaluations, the
better positioned selection boards will be to make
the difficult choices inherent in talent management of
a large pool of candidates.

A key point on cognitive bias is that it is uninten-
tional. Evaluating a person’s performance is undoubt-
edly complex. How much of performance is due to a
person’s talent versus the interactive effects from the
group? And how does their performance compare to
their peers who faced similar tasks but did so un-
der different conditions with different teammates?
Psychologist Daniel Kahneman shaped much of
what we understand about complex decision-making
with his insights on System 1 and System 2 thinking.
System 1 thinking normally guides our decisions as it
operates automatically and enables us to make most
decisions with little or no effort. When faced with
more complex tasks, System 2 thinking enables us to
focus our attention on more complex computations.
While we like to think we can put System 2 in control
when needed, Kahneman suggests that System 1 often
takes over in the face of complexity.

For instance, if asked what you think the president’s
popularity will be six months from now, what system
would you use? Kahneman claims this is a System 2
task since an accurate answer would require a person to
consider the events between present time and six months
in the future that would potentially affect the presi-
dent’s popularity and render judgment on the likelihood
of these events. Instead of performing these complex
calculations, we rely on System 1 thinking, which would
use the president’s current popularity to gauge what his
popularity will be six months from now.

A similar process unfolds for performance evalua-
tion. To complete the difficult task of assessing someone’s
performance, we use shortcuts that rely on information that is already stored in memory. The benefit of System 1 thinking is that it enables us to rely on intuition to perform such complex tasks, but the downside is that this process invites bias. Our System 1 thinking may succumb to the following five sources of bias when faced with the complexity of performance evaluation. The more we are aware of these biases, the better equipped we are to slow down our System 1 thinking and engage some System 2 functions to counter these biases.

**Halo effects.** As the name implies, halo effects occur when we use performance in one dimension to influence our evaluation of a person in all other dimensions. The primary problem of halo effects is that they decrease the number of opportunities for a person to demonstrate his proficiency, thereby precluding the rater from evaluating the ratee accurately across different dimensions of performance.\(^\text{21}\) Raters are especially susceptible to halo effects in systems where a single evaluator rates a person on multiple dimensions—as is the case with our evaluation system and the Army leadership requirements model with its core competencies and attributes.\(^\text{22}\)

The halo effect can be positive or negative. For example, an officer who performs well in the attribute of competence by projecting self-confidence and a commanding presence may enjoy a positive halo effect across the other competencies and attributes. Conversely, an officer who shows a lack of self-confidence and commanding presence may suffer a negative halo effect across the other competencies and attributes.

**First impression error.** This bias stems from initial impressions, either favorable or unfavorable, that influence a rater’s evaluation. Similar to halo effects, the primary problem of initial impression error is that a rater may suppress or discount subsequent information about a ratee if it is counter to their initial impression.\(^\text{23}\) This effect can be especially prevalent when a senior rater rates a large pool of a particular position or rank and has few interactions with each individual.

**Similar to me effect.** This bias stems from a tendency of some raters to judge a person favorably when he or she resembles the rater along dimensions such as his or her attitude or background.\(^\text{24}\) Some recent studies indicate that the military may be especially susceptible to this bias in comparison to other professions. A study of Army War College students found that this population scored lower on openness than the general U.S. population.\(^\text{25}\)

A characteristic of people with low scores on openness is that they prefer familiarity over novelty; thus, lower scores for openness may be associated with less favorable judgments of ratees who are significantly different than the raters. Other studies indicate service academy cadets scored lower on innovative cognitive style (which is positively correlated with a willingness to adopt new ideas) than students at comparable civilian universities, and those who left the academy after their first year scored higher on innovation than those who remained.\(^\text{26}\)

A study of the relationship between cognitive ability and promotion/selection found that officers with significantly higher cognitive abilities had 29 percent lower odds of selection below the zone (ahead of peers) to major, 18 percent lower odds for selection below the zone to lieutenant colonel, and 32 percent lower odds for selection to battalion command.\(^\text{27}\) One explanation for these results is that officers with high cognitive abilities may make “worse” junior officers since they may be less likely to be hypercompliant in comparison to those of average or lower cognitive ability. By this reasoning, the “similar to me effect” may contribute to these results.

**Central tendency error.** The central tendency error occurs when raters score most ratees as average or slightly above average.\(^\text{28}\) Although there are four blocks on the officer evaluation report, raters rarely use the “qualified” or “not qualified” box. While there are consequences for a rater to “bust their profile” by scoring too many officers as “most qualified,” there are no consequences for placing too many officers in the “highly qualified” category.

In situations where there are no consequences for too many average ratings, there is a greater potential for ratings inflation.\(^\text{29}\) Qualified or not qualified ratings involve additional work for the rater in terms of greater potential for interpersonal conflict with the ratee or the requirement for performance counseling documents if the rated officer appeals the evaluation. Since no consequences exist for establishing gradations in the quality of performance for those who are not “most qualified,” it is easier to rate someone as “highly qualified” than to use the lower two rankings. While our professional ethos is a check against this bias, we include it in this discussion since the potential exists for this bias.

**Duration neglect.** The essence of duration neglect is the tendency to place greater emphasis on peak time periods and recency when recalling events. To illustrate this effect, Kahneman discussed
a study of how patients recalled a colonoscopy. While the duration of the procedure had no effect on the patients’ ratings of total pain, the average level of pain at the worst moment of the procedure and at the end of the procedure were strong predictors of the overall evaluation of pain.

Hopefully, pain is not an emotion that raters recall during an evaluation, but the general principle applies for how this bias may influence evaluations. Instead of engaging System 2 processes to consider the performance of a ratee over a series of events, it is easier to use a key event such as an inspection, a training exercise, or the most recent training event to shape the impression a senior rater wishes to convey in an evaluation.

**Addressing Cognitive Biases**

We suggest three ways to counter these cognitive biases. Reading this article and becoming aware of countering sources of cognitive bias is the first step. While we hope that readers will find this information helpful, we think it is especially important to include education on these biases as part of professional military education. Professional military education courses often cover board processes and trends, but they do not currently include training on these biases. We think that just as future battalion and brigade commanders receive training on managing their profile, they should receive training on rater biases to become better evaluators.

Second, since the source of these biases is a system that relies on evaluations by a single rater, we recommend that raters seek input from different sources to help form their judgment of a ratee. One of the authors has experience with this technique while serving as a battalion executive officer. The battalion commander asked the operations officer, command sergeant major, senior chief warrant officer, and the author to rank the six company commanders. After submitting the feedback, the author compared his recommendations with those of the operations officer and found that his ratings were the opposite for the six commanders. While differences of opinion will probably not always be this stark, there is value in raters receiving a diversity of opinions to counter possible sources of cognitive bias.

Third, frequent feedback to subordinates can help counter bias, especially if a rater is aware of the potential biases discussed above. Frequent feedback can foster agreement on performance standards and increase acceptance of feedback by subordinates. This is an area that many leaders struggle with. In the 2016 Center for Army Leadership Annual Survey of Army Leadership, over one-third of respondents reported their supervisors rarely or never took time to discuss how they were doing with their work and what they could do to improve their performance.

**Conclusion**

In reality, the Army’s performance appraisal system is a multiyear assessment that is prone to disparities between senior raters and the profiles they maintain. As this article demonstrates, there are structural and cognitive biases that may affect the rating an officer receives. These biases undermine the meritocratic principles that we seek in our performance evaluation system. The more that we are aware of these biases, the better position we will be in to counter their effects.

Editor’s note: We wish to express our appreciation to library research archivists Russell Rafferty and Elizabeth Dubuisson of the Ike Skelton Combined Arms Research Library, Fort Leavenworth, Kansas, for their support in locating early versions of Army efficiency reports and references to them in period official technical manuals.

**Notes**


13. Ibid.


16. Ibid.


29. Murphy and Cleveland, Understanding Performance Appraisal.

30. Ibid.