2024-06

FITREP CONTENT RELIABILITY: AN INTER-RATER PERSPECTIVE

Banks, Fatima N.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

FITREP CONTENT RELIABILITY: AN INTER-RATER PERSPECTIVE

by

Fatima N. Banks

June 2024

Thesis Advisor: Chad W. Seagren
Second Reader: Mitchell Friedman

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704–0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2024	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE FITREP CONTENT RELIABILITY: AN INTER-RATER PERSPECTIVE			**5. FUNDING NUMBERS**
6. AUTHOR(S) Fatima N. Banks			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A. Approved for public release: Distribution is unlimited.			**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The reliability of the Performance Evaluation System within the Marine Corps has come under scrutiny, particularly regarding the consistency and fairness of fitness report evaluations. This study aimed to investigate inter-rater reliability among Reporting Seniors (RSs) in assessing the performance of Marine Officers using the Performance-Anchored Rating Scale (PARS). Data collected from 51 respondents were analyzed, revealing notable findings. Quantitatively, the study found low inter-rater reliability, with Fleiss' Kappa yielding a value of 0.0214 ($p = 0.00116$), indicating inconsistent application of assessment criteria across diverse backgrounds. These results raise concerns regarding the validity and fairness of fitness report evaluations and their ability to accurately identify talent and quality Marines for promotion within the Marine Corps. Despite efforts to standardize evaluation criteria and promote objectivity, the analysis suggests substantial variability in ratings and evaluations among RSs. Recommendations stemming from these quantitative findings include reassessing evaluation criteria and considering modifications to the PARS and evaluated attributes to enhance accuracy and fairness. Addressing these issues is imperative for ensuring fairness, transparency, and equity in personnel management practices within the Marine Corps.

14. SUBJECT TERMS USMC, Marine Corps, performance evaluation, fitness report, FitRep, reliability, PES, Performance Evaluation System			**15. NUMBER OF PAGES** 123
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

FITREP CONTENT RELIABILITY: AN INTER-RATER PERSPECTIVE

Fatima N. Banks
Captain, United States Marine Corps
BA, Southern New Hampshire University, 2017

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
June 2024

Approved by: Chad W. Seagren
Advisor

Mitchell Friedman
Second Reader

Marigee Bacolod
Academic Associate, Department of Defense Management

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The reliability of the Performance Evaluation System within the Marine Corps has come under scrutiny, particularly regarding the consistency and fairness of fitness report evaluations. This study aimed to investigate inter-rater reliability among Reporting Seniors (RSs) in assessing the performance of Marine Officers using the Performance-Anchored Rating Scale (PARS). Data collected from 51 respondents were analyzed, revealing notable findings. Quantitatively, the study found low inter-rater reliability, with Fleiss' Kappa yielding a value of 0.0214 ($p = 0.00116$), indicating inconsistent application of assessment criteria across diverse backgrounds. These results raise concerns regarding the validity and fairness of fitness report evaluations and their ability to accurately identify talent and quality Marines for promotion within the Marine Corps. Despite efforts to standardize evaluation criteria and promote objectivity, the analysis suggests substantial variability in ratings and evaluations among RSs. Recommendations stemming from these quantitative findings include reassessing evaluation criteria and considering modifications to the PARS and evaluated attributes to enhance accuracy and fairness. Addressing these issues is imperative for ensuring fairness, transparency, and equity in personnel management practices within the Marine Corps.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

APES	Automated Performance Evaluation System
CA	comparative assessment
CNA	Center for Naval Analysis
DOR	Date of Rank
EWS	Expeditionary Warfare School
FitRep	fitness report
FY	fiscal year
HQMC	Headquarters Marine Corps
iAPS	Improved Awards Processing System
MBS	Master Brief Sheet
MCO	Marine Corps Order
MCTFS	Marine Corps Total Force System
MMMA	Manpower Management Military Awards
MMRP	Manpower Management Records and Performance Branch
MOS	military occupational specialty
MRO	Marine Reported-on
NAVMC	Navy and Marine corps
NPS	Naval Postgraduate School
PARS	Performance-Anchored Rating Scales
PES	Performance Evaluation System
OMPF	Official Military Personnel File
RCLF	Regional, Culture and Language Familiarization
RO	Reviewing Officer
RS	Reporting Senior
RV	relative value
SOP	Standard Operating Procedures
STEM	Science, Technology, Engineering, and Mathematics
TBS	The Basic School
TIG	Time-in-Grade

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I extend my deepest gratitude to my friends and family for their unwavering support during my journey at the Naval Postgraduate School. Through every triumph and challenge, their presence has been a beacon of strength, keeping me motivated and inspired.

A heartfelt thank you goes out to my son, Roman, whose resilience, understanding, and adaptability made the transition to Monterey seamless. His support has been a driving force behind my pursuit of success, and for that, I am profoundly grateful.

I am also indebted to my esteemed advisors, Dr. Chad Seagren and Dr. Mitchell "Can you cite this?" Friedman. Dr. Friedman played a pivotal role in refining my thought process and ensuring that my ideas were effectively communicated in my paper. Meanwhile, Dr. Seagren recognized the potential in my concepts and provided invaluable guidance in refining my analysis and steering me back on course whenever I allowed personal biases to influence the direction of my work.

To each and every one of you who has contributed to my academic journey, I offer my sincerest thanks. Your support, guidance, and encouragement have been instrumental in my growth and achievements.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This study investigates the reliability of Marine Corps fitness report (FitRep) evaluations, a pivotal aspect of the Performance Evaluation System (PES). Within the broader context of organizational and human resource management, the study focuses on addressing unexplained scoring inconsistencies through the lens of statistical reliability.

Empirical literature reviews and peer discussions suggest that the current FitRep system may lack the requisite structure for a comprehensive, fair, and unbiased assessment of talent and performance. Challenges such as trait selection ambiguity, validity concerns, and the FitRep's multifaceted utility have cast doubts on its efficacy. Moreover, the PES grapples with inadequate structuring, vague performance assessment criteria, and a reliance on subjective Reporting Senior (RS) opinions. Anecdotal evidence suggests that rather than adhering to established criteria, RSs often use their grading profile, leading to scores reflecting more on how a Marine compares to past evaluations by the same RS than on the Marine's actual performance.

Concerns about unexplained score variations based on Marine Reported-On (MRO) race, gender, and Military Occupational Specialty (MOS) were highlighted in a major study conducted by the Center for Naval Analysis (CNA) (Clemens et al., 2012). Multiple studies conducted by Naval Postgraduate School (NPS) students, including Jobst and Palmer (2005), Rigaut (2017), and Larger (2017), found similar issues—as did Dunst (2018) who also found disparities in scores based on college degree program in addition to the aforementioned demographics.

After-action reports from selection board members, such as that from Heuer (2020), echoed concerns about the quality of information, contradictory scoring, and inconsistent word pictures in evaluations. These factors make it challenging to obtain an accurate and complete understanding of an MRO when determining suitability for promotion and special programs (Heuer, 2020).

Additionally, the 38th Commandant's Planning Guidance acknowledged major shortcomings in the existing PES, recognizing a "growing lack of faith within our ranks in

the system's ability to accurately identify their skills, performance, and future potential" (Berger, 2019). However, despite prioritizing modifications, no funding has been allocated to evaluate proposed changes.

This study offers a fresh perspective in its efforts to understand FitRep impartiality by investigating reliability as a potential contributing factor. We have two research questions:

1. Primary Question: To what extent does the inter-rater reliability of FitRep evaluations among RSs validate the assessment criteria used in measuring performance?

2. Secondary Question: What factors or characteristics of RSs correlate with trends in FitRep scores and inter-rater reliability?

Together, these questions explore the extent of RS consistency in ratings and evaluations in Marine Corps FitReps and further explore factors contributing to potential variations in reliability. We hypothesize that there is low inter-rater reliability among RSs in their evaluations of FitReps, suggesting that the assessment criteria used in measuring performance are not consistently applied across raters and do not validate performance assessments. Focused on Marine Corps officers, the study employs a survey mimicking sections D through G of the current Navy and Marine Corps (NAVMC) 10835 (Rev. 7–11) FitRep to assess the content reliability of the FitRep.

Our findings reveal that the inter-rater reliability of FitRep evaluations among RSs does not fully validate the assessment criteria for measuring performance. Despite efforts to ensure consistency and fairness, discrepancies in ratings and interpretations persist among RSs. Additionally, the lack of statistically significant differences in inter-rater reliability across various training groups suggests that factors beyond formal training influence variations in FitRep evaluations.

While definitive correlations between specific RS characteristics and FitRep trends are lacking, our analysis highlights the complex interplay of factors impacting rating consistency and reliability. We find that officers at higher ranks or with specific

occupational specialties may have differing perspectives or expectations regarding performance, leading to rating variations. Qualitative insights from our text analysis underscore the importance of contextual knowledge and direct observation in evaluations, suggesting that subjective aspects often outweigh the intended objectivity of attribute descriptions and PARS ratings. Furthermore, inconsistencies between narrative justifications and PARS ratings, particularly for Marines ranked as average overall, indicate potential biases or subjective interpretations influenced by individual characteristics such as experience level, communication skills, and critical thinking ability.

Given the critical shortcomings in the current framework, research into the Marine Corps PES is imperative. The system's lack of comprehensiveness in defining quality standards, measures for special attributes and skills, and alignment with organizational needs underscores the need for systematic evaluation. Marine Corps Order (MCO) 1610.7B serves as a guide for completing performance evaluations using the FitRep, yet challenges persist in subjective factors, space constraints for comments, and MOS-specific trait assessment.

These issues undervalue crucial skills that optimize performance or value but are not identified as a particular billet responsibility or requirement. By addressing reliability concerns, this study aims to furnish quantitative evidence supporting these concerns, informing strategic decisions, and ensuring the integrity of promotion and retention processes within the Marine Corps.

The subsequent chapters of this thesis are organized as follows:

1. Chapter II presents the definition and value of statistical reliability tests in examining performance measures, an overview of the PES and its structural and complementary components.

2. Chapter III consists of a comprehensive review of literature on performance evaluations including studies seeking to identify commonalities in effective performance evaluations and understand and validate measurement tools. This discussion is followed by a review of studies on the Marine Corps PES and the issues and challenges identified.

3. Chapter IV outlines methods employed to source data and analyze results.

4. Chapter V contains the analysis.

5. Chapter VI concludes the study and presents discussion and recommendations for change.

II. BACKGROUND

This chapter provides a comprehensive overview of the Marine Corps PES. It navigates through the PES components, including objectives, history, and roles and responsibilities.

Complementary key documents, such as the Master Brief Sheet (MBS) and Official Military Personnel File (OMPF), are introduced to illustrate their role in presenting a consolidated view of a Marine's service history and performance records. Finally, drawing on insights from a board member for a fiscal year (FY) 2019 promotion board, the chapter discusses the FitRep's implications for? promotion boards and challenges in conveying meaningful information.

A. USMC PERFORMANCE EVALUATION SYSTEM

The PES supports personnel management and assignment decisions including unifying selection, promotion, and retention efforts across both Active and Reserve components (Headquarters, United States Marine Corps, 2023). The PES Manual establishes foundational guidelines for fitness reports, encompassing policies, procedures, and responsibilities. The subsequent details on system objectives, evaluation roles and responsibilities, and NAVMC 10385 description are sourced from this manual.

1. System Objectives

The system objectives, outlined in the PES Manual, aim to achieve accurate fitness reports that assess individual performance against specific criteria (Headquarters, United States Marine Corps, 2023). This intention highlights the importance of clarity and objectivity, aligning with Marine Corps standards to avoid inflated performance assessments. Timeliness is vital, with normal reports expected within 30 days and adverse reports within 60 days. This emphasis on timeliness ensures prompt and precise updates to Marines' Official Military Personnel Files (OMPFs) (Headquarters, United States Marine Corps, 2023).

The PES Manual encourages the RS to submit fitness reports that are both administratively and procedurally correct (Headquarters, United States Marine Corps, 2023). Doing so ensures that there are accurate and complete records of each Marine's performance over time (Headquarters, United States Marine Corps, 2023). This emphasis on accuracy and adherence to guidelines helps in maintaining comprehensive and continuous documentation of the performance of every Marine within the system. It also supports selection boards by providing fair and accurate information for personnel management decisions (Headquarters, United States Marine Corps, 2023).

2. History of the Performance Evaluation System

In 1999, the Marine Corps implemented the current FitRep system to address grade inflation, as detailed in Phillips and Clemens (2011). Several modifications were introduced. The changes included reducing evaluation dimensions from 21 to 14 and introducing text boxes for billet description and accomplishments (Phillips & Clemens, 2011). Additionally, the previous FitRep had a 6-point rating scale, while the current system employs 14 Personal Anchored Rating Scale (PARS) ranging from A to G (with H indicating not observed) (Phillips & Clemens, 2011). Notably, an A in any PARS signifies an "adverse" report, with F and G requiring justification (Phillips & Clemens, 2011).

Another significant disparity is that the current FitRep calculates an unweighted average of the PARS, eliminating the overall RS mark (Phillips & Clemens, 2011). The MRO is assigned a relative value (RV) based on the RS's profile, reflecting a comparative analysis with other MROs with the same RS and grade.

Lastly, unlike the previous system, which lacked the ability to track of reporting history, the new FitRep introduces an overall relative assessment by the Reviewing Officer (RO), aiming for a distribution resembling a "Christmas tree" and replacing the previous absence of a numerical RO mark (Phillips & Clemens, 2011). These changes, along with others, contributed to the expansion of the evaluation format to the current 5-page document, compared to the previous 2-page format (Phillips & Clemens, 2011).

Over the years, the PES has received minor upgrades to account for changes in requirements such as the inclusion of the combat fitness test (CFT) as an annual training requirement.

3. Roles and Responsibilities

Roles and responsibilities, outlined in the PES Manual, designate the MRO as the focus of the report, required to route accomplishments to the RS via the Marine Reported-On Worksheet (MROW). See Appendix A for an example of the MROW. The RS, typically the first commissioned or warrant officer senior to the MRO, is the direct supervisor. He or she asses the performance and character of a MRO using the NAVMC 10835 FitRep, and routes to the Manpower Management Records and Performance Branch of HQMC (MMRP-30) via the Automated Performance Evaluation System (A-PES) (Headquarters, United States Marine Corps, 2023). The Reviewing Officer (RO), senior to the RS, is the direct supervisor to the RS (Headquarters, United States Marine Corps, 2023). Manpower Management Records and Performance Branch (MMRP-30) (Headquarters, United States Marine Corps, 2023).

a. RS Marking Philosophy

Marking philosophies are unique to individual RSs and are developed based on PARS descriptors. The marking philosophy, integral to the evaluation process, serves two primary purposes: providing RSs with a method to assess Marines across attributes and establishing a practical scale for communicating expectations (Headquarters, United States Marine Corps, 2023). Once established, RSs must maintain consistent marking philosophies across all ranks and MOSs throughout their careers to eliminate subjectivity and ensure fair evaluations (Headquarters, United States Marine Corps, 2023). Any mid-career changes to marking philosophies could have significant effects on previously evaluated Marines, emphasizing the RS's responsibility to minimize subjectivity by adhering to objective criteria (Headquarters, United States Marine Corps, 2023).

When completing a FitRep, RSs concentrate on delineating the Marine's actions and achievements during the specific reporting period. Post-report completion, the RS compares assigned attribute marks with those of other FitReps they have written on MROs

in the same grade. This process enables minor adjustments to maintain consistency with an established marking philosophy and align with the RS's historical standards (Headquarters, United States Marine Corps, 2023).

b. Reporting Senior Profile

The FitRep incorporates both an absolute and relative measure of assessment in the form of personal attributes with PARS and the RS Profile, respectively. Both measures are integral components of the PES. PARSs, discussed later in this chapter, offer comprehensive descriptors for each assessed attribute, including clear definitions, descriptions of performance levels, and a marking gradient (Headquarters, United States Marine Corps, 2023). The RS Profile encompasses the average FitRep scores, arranged in descending order, for all reports crafted by an RS (Headquarters, United States Marine Corps, 2023). An RS generates a "profile" after he or she has evaluated three MROs in the same grade. After evaluating three MROs in the same grade, the RS generates a "profile," enabling them to review and maintain consistency with past markings, as well as to establish a relative ranking of MROs within the same grade (Headquarters, United States Marine Corps, 2023).

4. NAVMC 10835 Fitness Report

In essence, the PES Manual guides the RS in completion and submission of a FitRep. See Appendix B for a blank example of the NAVMC 10835 FitRep. The FitRep is instrumental in determining eligibility and suitability for promotion, shaping career paths, and aiding in retention decisions. Its pivotal role extends beyond a simple evaluation to a critical piece in support of the Commandant's overarching objectives (Headquarters, United States Marine Corps, 2023). To provide a comprehensive understanding of the NAVMC form, this section explains the various sections and their significance, as outlined in the PES Manual.

a. Administrative Information: Section A

Section A of the fitness report serves the purpose of providing essential administrative information about the MRO. It collects critical administrative details about

the Marine, the reporting organization, the occasion, and period covered, duty assignment, and any special cases (e.g., adverse, or commendatory material). Section A establishes a foundational framework for the subsequent sections of the fitness report.

b. Billet Responsibilities: Section B

Section B of the FitRep provides the RS with a chance to outline the range of responsibilities that served as the foundation for the evaluation during the specified reporting period. (Headquarters, United States Marine Corps, 2023). The PES Manual provides details to guide the development of the billet description including:

1. Avoiding restating MOS prerequisites, and
2. Emphasizing both the MRO's role and their primary responsibilities within the unit or organization.

Due to space limitations, the billet description should emphasize the most relevant aspects of the billet, prioritizing acceptable standards over goals (Headquarters, United States Marine Corps, 2023).

c. Billet Accomplishments: Section C

Section C focuses on highlighting the MRO accomplishments deemed most significant by the RS for the reporting period (Headquarters, United States Marine Corps, 2023). It is completed by the MRO and has space to include course completion, class standings, and other relevant achievements. The MRO provides a comprehensive account of his or her accomplishments in the billet, aligning with the information in Section B. The PES Manual instructs the MRO to present only factual results and achievements avoiding any reference to the influence of individual merit, maintaining a concise and focused presentation of performance.

d. Evaluation: Sections D, E, F, G, and H

Sections D through H of the FitRep encompass 14 attributes crucial for evaluating the Marine's performance. See Table 1 for a comprehensive list and description of all attributes. These attributes, outlined in PARS, offer comprehensive descriptors for each

evaluated attribute, including definitions, performance levels, and marking gradients. PARS aim to streamline the evaluation process, reducing the need for extensive written comments while promoting objectivity and consistency based on Marine Corps expectations (Headquarters, United States Marine Corps, 2023). The markings range from "A" to "H" and correspond to scaled descriptions. They are designed to stimulate critical thinking and thoughtful analysis, helping the RS evaluate the MRO more effectively before making a decision (Headquarters, United States Marine Corps, 2023).

Table 1. NAVMC 10835 Fitness Report attributes and definitions. Source: Headquarters, United States Marine Corps (2023)

Section	Description	Definition	Attribute
D	Mission Accomplishment	Addresses both the ends (results) and the means (how the MRO achieved those results).	Performance
			Proficiency
E	Individual Character	Focuses on measurable traits of the MRO's individual character such as distinctive mental, physical, moral, and behavioral qualities that each Marine needs.	Courage
			Effectiveness Under Stress
			Initiative
F	Leadership	Provides overall view and understanding of the individual's leadership style, in addition to comprehensive picture of the individual's effectiveness as a leader.	Leading Subordinates
			Developing Subordinates
			Setting the Example
			Ensuring Well-Being of Subordinates
			Communication Skills
G	Intellect and Wisdom	Measures the MRO's efforts to grow intellectually, and use knowledge gained to benefit both personal and unit performance. Provides a critical indicator of an MRO's ability to learn and reason, capacity for knowledge and understanding, and ability to use intellectual skills to make viable and timely decisions.	Professional Military Education
			Decision Making Ability
			Judgment
H	Fulfillment of Evaluation Responsibilities	Establishes a direct method of ensuring that reporting officials accomplish the objectives of the PES by evaluating their efforts to submit accurate, timely, and uninflated evaluations. Measures the level to which reporting officials fulfill their responsibilities.	

5. Master Brief Sheet

The MBS consolidates personal information and summarizes the performance evaluation record. Comprising two distinct sections, the Header Data extracts service information directly from the Marine Corps Total Force System (MCTFS), providing essential details about the Marine's service history (Headquarters, United States Marine Corps, 2023). On the other hand, the Fitness Report Listing section offers an overview of archived reports highlighting sections A through H, and a portion of section K (Headquarters, United States Marine Corps, 2023). Essentially, the MBS serves as a centralized repository, streamlining access to critical information for effective personnel management.

B. OFFICIAL MILITARY PERSONNEL FILE

The OMPF serves as a comprehensive repository, capturing the entirety of a Marine's military career, including administrative details, awards, MOS, training summaries, languages, education summaries, official photos, and fitness reports (Headquarters, United States Marine Corps, 2023). It is imperative for the MRO to meticulously verify all information as accurate and current, ensuring that all documentation, including training certificates and official photos, is accounted for.

C. FITNESS REPORT CHALLENGES IN PROMOTION BOARD DECISION-MAKING

After establishing the foundation by emphasizing the significance of the FitRep within the PES and recognizing the current challenges in ensuring an equitable assessment of talent and performance, the next step is to integrate these considerations with the valuable perspectives shared by Lieutenant Colonel Jason W. Heuer (2020). As a board member for a FY19 promotion board, Heuer's insights further illustrate the challenges and intricacies associated with the FitRep in the context of the broader goal of improving fairness and accuracy in personnel assessments.

Heuer's (2020) after-action report highlights the crucial role of FitReps in the promotion process. He emphasizes the board's commitment to thoroughly evaluating each document and stresses the importance of effective communication through FitReps.

Heuer (2020) notes that the board scrutinizes reports meticulously and expects them to be written to the board rather than the Marine.

Common issues include consecutive low marks without adequate explanation, inconsistencies in attribute marks between the RS and RO, and reliance on cliché phrases (such as "completed tasks with minimal supervision," and "valued member/asset of the command") rather than language that better illuminates the rationale for specific marks (Heuer, 2020, p. 50). These observations highlight challenges associated with the current PES, aligning with our research goal of investigating reliability concerns.

Heuer's (2020) after-action report provides valuable insights that support our research goal by highlighting the meticulous scrutiny given to Fitness Reports during promotion boards and the need for improvement in conveying meaningful information for fair and accurate personnel assessments.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LITERATURE REVIEW

This chapter offers a thorough exploration of research on effective performance evaluation measures, encompassing insights into common traits and characteristics, tests of validity and reliability, and the challenges specific to the Marine Corps PES. These themes, inclusive of civilian and military organization perspectives, collectively enhance our understanding of the complexities surrounding performance evaluations and assessment measures, setting the stage for why statistical reliability becomes the next crucial step in evaluating the Marine Corps PES.

A. EFFECTIVE PERFORMANCE EVALUATION MEASURES

A thorough performance evaluation incorporates several elements to ensure a comprehensive and impartial assessment of an individual's work. While the specific factors might differ based on the context and nature of the work, there exist some commonalities found in different organizations in various industries which form an integral part of effective performance measures.

1. Common Traits and Characteristics

Chachula (1992) attempted to identify such commonalities by examining the performance measurement systems of 11 successful companies. She used field reports compiled from interviews with the companies' leadership, including site supervisors, managers, controllers, quality managers, and individuals in roles such as purchasing, maintenance, or product design, to identify 21 "best practices" in performance evaluations by corporate standards (Chachula, 1992). The data originated from prior research, conducted by Euske, Lebas, and McNair (1993), obtained via site visits to 11 commercial companies located throughout the U.S. and Europe. Interview questions were designed by the researchers to elicit direct and indirect information about the development, utility and perception of performance measures used by the companies (Chachula, 1992).

Chachula (1992) employed data analysis techniques such as data reduction, display, and conclusion drawing to identify theoretical characteristics of the companies'

performance measures from the compiled reports. The results initially identified 37 common characteristics, which were then condensed to 21 based on similar themes.

The examination of performance measures highlighted key trends across management levels. See Table 2 for summary of characteristics and their significance to management. Notably, there was a consistent preference for physical measures over financial ones, with financial metrics gaining importance in challenging market conditions and for top-level management. Teamwork and the value of self-directed work teams emerged as crucial factors, echoing across different management tiers. Evaluation processes were well-established, with semi-annual or annual reviews being common, supplemented by informal evaluations. People skills, time frame considerations, and the universal recognition of the workforce's significance were consistent themes.

Concerns about supplier relationships, acknowledgment of technology's impact, and the focus on quality were noteworthy aspects, each with variations across management levels. Compensation patterns, communication strategies, customer focus, and empowerment were consistent themes across organizations and management levels, each playing a pivotal role in organizational effectiveness. The study also highlighted the emphasis on leadership and management training, aligning with a broader focus on physical performance measures. Continuous improvement and professional growth training remained pervasive considerations at every management level, along with the enduring importance of key measures of production and productivity.

Chachula (1992) considered each characteristic foundational for effective performance measures, influencing aspects such as promotions, performance evaluations, compensation, and office climate. The identified characteristics were analyzed to establish their connection to the reward structures within the organizations. This analysis revealed a key insight: in all 11 companies studied, every characteristic, except for supplier relationship, time frame, and technology, received rewards through one or more of the influential aspects mentioned.

Table 2. Summary of organizational performance measures and key themes across management levels. Source: Chachula (1992)

Common Characteristics of Performance Measurement Systems		
Physical/Financial Measures: Most sites prioritized physical over financial measures, except in challenging market conditions. Financial measures were crucial for top-level management.	Teamwork: Teamwork and the value of self-directed work teams were crucial for successful organizations.	Cycle Time: Mentioned in manufacturing organizations, cycle time was consistently discussed across management levels.
Changes in Performance Measures: Despite initiatives like self-directed work teams, changes in performance measures predominantly originated from top management.	Community Involvement: While not universal, involvement with the local community was notable in the top two levels of management.	Market Related: PMP Managers were particularly concerned about market-related aspects.
Evaluations: All organizations conducted established performance evaluations semi-annually or annually, with informal evaluations used across all management levels.	People Skills: People skills were highlighted by at least one manager in every organization.	Time Frame: Time frame considerations were evident in organizations with a focus on long-term contracts.
Importance of Employees: Employees were universally recognized as a valuable resource across all sites and management levels.	Supplier Relationship: Concerns about supplier relationships were expressed by Site Managers, Quality Managers, and PMP Managers.	Technology: Most organizations acknowledged the impact of technology, with variations in concerns across management levels.
Quality: Quality was a central theme guiding organizations and management at every level, with variations in its interpretation.	Compensation: Compensation patterns were consistent, primarily comprising base salary and performance-related bonuses.	Communication: Informal and formal communication were crucial in all organizations, with top management focusing more on informal communication.

Common Characteristics of Performance Measurement Systems		
Customer Focus: Customer focus was consistent across all organizations and management levels, emphasizing meeting customer needs.	Empowerment: Empowerment was recognized across management levels, with PMP Managers placing a higher emphasis.	Leadership/Management: Emphasis on leadership and management training aligned with a focus on physical performance measures.
Continuous Improvement: Emphasis on continuous improvement was evident at every management level, indicating a commitment to enhancing operations.	Training: Professional growth training was deemed important at every management level.	Productivity/Performance: Key measures of production and productivity were consistently important across all management levels.

Building upon Chachula's foundational exploration of effective performance evaluation measures, Small (2020) further contributes to our understanding of best practices in organizational success. In a more contemporary context, Small's (2020) research focused on high-performing companies like Google and Deloitte, exploring intricacies of modernizing employee performance evaluation systems. Small's empirical review resulted in four narrative themes capturing successful performance evaluation practices. She used these themes to structure a theoretical framework to help better inform the modernization effort of the Navy's personnel system using evidence-based research in scientific literature and industry practices (Small, 2020).

To comprehend the regulations and procedures governing performance evaluations, Small (2020) conducted a thorough review of foundational statutes and Navy publications. Additionally, she conducted a comprehensive search in academic databases, including EBSCOhost, JSTOR, Web of Science, Google Scholar, and the Dudley Knox Library, focusing on topics such as employee performance evaluation, appraisal, measurement methods, talent management, and evaluation methods.

Small (2020) used the same databases to identify Google and Deloitte by focusing her search on high-performing companies with over 100,000 employees that had

implemented transformative measures to modernize their employee PES between 2010 and 2020. The successful practices of performance evaluations are encapsulated in four narrative themes, addressing the key aspects of evaluation purpose, employee involvement, methods, rating considerations, and the evaluation process.

1. Clearly Define the Evaluation Purpose: This theme emphasizes the importance of identifying specific objectives and goals for performance evaluations. Small's (2020) research highlights challenges faced by the Navy PES due to competing objectives and the use of a single system for both developmental and administrative goals.

2. Cultivate a Culture of Communication: This theme focuses on establishing an environment that encourages open communication, feedback, and ongoing dialogue between employees and managers. Small (2020) identifies challenges in the Navy PES, such as minimal opportunities for feedback, a focus on process over performance, and a lack of transparency. Recommendations include the use of the MSAF coaching tool to enhance feedback and communication, simplifying the evaluation system, and fostering transparency through tools like 360-degree feedback.

3. Promote a Perception of Fairness: This theme underscores the importance of ensuring that employees perceive the evaluation system as fair and accurate. Small (2020) identifies challenges with accuracy in the Navy PES, leading to potential talent mismanagement. Recommendations include constant calibration and clear communication to enhance fairness, as well as simplifying processes and fostering transparent communication to address opaque processes and build trust.

4. Acknowledge the Absence of a Universal Solution: This theme recognizes the complexity of performance management systems and emphasizes the need for customized solutions based on organizational needs. Small (2020) concludes that no universal solution exists; effective implementation

requires a tailored approach, rigorous testing, and periodic audits. Successful PES implementation depends on integrating evaluation into overall performance management, iterative testing, and periodic audits.

Chachula (1992) and Small (2020)'s findings both contribute significantly to the understanding of effective performance evaluation measures. Both studies focus on defining evaluation purposes, emphasizing the importance of identifying specific objectives and goals for performance evaluations while ensuring accuracy aligns with the need for meaningful financial measures. Both emphasize the importance of teamwork and fostering a culture of communication as crucial to the success of an organization. Chachula (1992) found continuous improvement was evident at every management level, indicating a commitment to enhancing operations. Similarly, Smalls (2020) acknowledges the absence of a universal solution, emphasizing the need for customized solutions based on organizational needs, indicating a continuous improvement mindset. By combining Chachula's foundational insights with Smalls' contemporary perspectives, we emerge with principles potentially able to address modern challenges.

2. Relative versus Absolute Measures

Performance assessments can follow a relative or absolute format. While the relative approach involves comparing individuals against their peers, guiding decisions such as promotions and wages, absolute evaluations establish a standard without peer comparison (Gomez-Mejia et al., 2016). The latter provides specific feedback, fostering fairness in comparisons across departments and contributing to the clarity of performance standards (Gomez-Mejia et al., 2016). On the contrary, relative measures pose challenges in accurately discerning individual variations and introduce ambiguity about the absolute standing of an individual's performance (Gomez-Mejia et al., 2016).

Building on this understanding, Balakrishnan and Sivaramakrishnan (2016) explored the efficacy of absolute and relative performance evaluation systems. This investigation specifically focused on the absolute system and a tournament ranking system, exploring the implications of individual performance manipulation. The study employed a task assignment model with risk-neutral agents and a risk-neutral principle to assess the

relative screening efficacies of the two systems in the presence of performance manipulation.

The task assignment model involved a risk-neutral principal assigning two risk-neutral agents to two tasks, considering the fit between agent types and tasks (Balakrishnan et al., 2016). The principal aimed to minimize the expected loss from task assignment, with defined costs for type I and type II errors (Balakrishnan et al., 2016). The absolute system categorized an agent's performance by creating a binary signal (high or low) dependent on whether their actual performance surpassed a predetermined threshold (Balakrishnan et al., 2016). The ranking system ranked the agents giving the agent with the higher signal the better ranking (Balakrishnan et al., 2016).

Balakrishnan et al. (2016) found that, in the absence of performance manipulation, both absolute evaluations and ranking systems are preferred over a no-information benchmark. Absolute evaluations optimally assigned tasks based on signals, considering higher signals for task 1 and lower signals for task 2, while ranking systems used ranking to make task assignments leaving the principal to decides whether to assign task 1 to the agent ranked first or to the agent ranked second.

Moreover, determining the optimal task assignment became inconsequential when both agents shared the same talent status, be it talented or untalented (Balakrishnan et al., 2016). In such instances, random assignment is invoked as a strategy to minimize the principal's loss. Additionally, Balakrishnan et al. (2016) suggested that the impact of information discretization within the absolute system diminishes with increased flexibility in task assignments. They recommended expanding the scope to consider numerous agents to enhance the appeal of absolute over ranking systems. Balakrishnan et al. also advocate for the exploration of multiple performance ratings, extending beyond the binary categorization of high and low.

Balakrishnan et al.'s (2016) findings are limited in that they do not realistically account for on-hand agents and the potential presence of indistinguishable talent. Researchers do not speak to the estimated effects in a practical application of this model within a company with tens or more employees. Most businesses and organizations have

more than two employees; additionally, some tasks may be linked to a position rather than flexibly distributed among teams. In such cases, it is unclear whether the conclusion still holds true regarding the preference for absolute measures. Moreover, employing multiple employees is likely to result in several agents receiving the same rating. In these instances, when the talent pool includes both standout agents and some who are indistinguishable, it is unclear how the measurement model would compensate.

3. Content Validity and Accuracy

Content validity, assessing the alignment of empirical measurements with a specific content domain, is pivotal in ensuring that a set of sample items effectively represents and adequately defines the measured construct (Carmines & Zeller, 1979; Haynes, Richard, & Kubany, 1995). However, challenges emerge when dealing with unobservable constructs that measure abstract or complex concepts, such as "leadership effectiveness," which involves multifaceted skills and behaviors. The subjectivity inherent in such constructs requires judgment from subject matter experts to assess content relevance and representativeness (Shrotryia & Dhanda, 2019).

Addressing the nuanced challenges of content validity, particularly when grappling with abstract constructs, Luke (2022) navigated these complexities by applying a robust examination to the U.S. Navy's trait value statements (TVS) in draft. Using cross-textual analysis between the TVS and authoritative doctrine and policies, Luke (2022) found high construct validity based on the correlation between the traits, doctrine and evaluations used by other military services.

As part of a larger overhaul of the personnel management system, the TVS included 8 traits, with various sub-traits and justifying statements all of which required assessment to determine their alignment with Navy doctrine (Luke, 2022). To better structure the FitRep into a coaching tool, "the language in the TVS was purposely developmental and included many parallels to civilian research on talent management" (Luke, 2022, p. 45).

Luke employed cross-textual analysis to assess the construct validity of the Navy's TVS with organizational goals and values identified in the following sources: Task Force One–Navy (TF1N) report, the Hard Truths and the Duty to Change: from the Independent

Review Commission on Sexual Assault report, The Chief of Naval Operations (CNO) 2021 NAVPLAN, the CNO's 2020 Signature Behaviors of the 21st Century Sailor document in addition to values outlined in other U.S. DOD military officer evaluations as identified in their PES. The analysis involved determining the frequency of each sub-trait's reference in these documents and ranking them from most to least valid based on their prevalence in all source documents (Luke, 2022).

Out of the 39 TVS sub-traits, 33 were validated by at least half of the Navy doctrine under review (Luke, 2022). 18 TVS sub-traits found exact matches in all three other services' officer evaluations while wellness was found to align with all reviewed Navy doctrine and evaluations from other services (Luke, 2022). The TVS sub-traits that received the highest rankings, meaning validation from six or more sources, primarily center around interpersonal dynamics, encompassing wellness, ethics, personnel development, innovation, relationships, feedback, inclusion, integrity, professionalism, feedback, personal development and listening (Luke, 2022).

The proposed new traits, emphasizing "leadership skills" with minimal connection to performance or proficiency, align with Small's (2020) conclusion regarding the lack of a clear objective within the Navy PES. Small highlighted challenges arising from competing objectives and the utilization of a single system for both developmental and administrative goals, reinforcing the need for clarity in defining evaluation purposes. Luke (2022) suggests a further misalignment of PES goals and utilization by drafting the TVS with a language geared toward coaching while still using the FitRep to evaluate performance and make decisions regarding promotion.

4. Reliability and Dependability

Test reliability, often measured by the correlation between scores for a representative group, provides a valuable index of the consistency and stability of the measurement over repeated assessments. To examine reliability and investigate the importance and relevance to performance evaluation measures, Bottoms (2022) conducted a study focused on evaluating the reliability of job performance test scorings through interrater consistency.

The analysis of 57 evaluators from the Early Education Support Office (EEOS) on 10 fictitious teacher profiles found varied degrees of evidence of strictness, leniency, and bias in the evaluation of early educators. Findings suggested inconsistencies in methods affecting birth-through kindergarten teacher licensure for early educators in both private and public sectors. Bottoms (2022) used the Strictness Calibration, Bias Test, and Caseloads Added models to investigate strictness and leniency patterns.

The evaluators, who hold licenses themselves, each rated 10 fictitious teacher profiles crafted from real cases, to assess model comparisons and goodness of fit using the 30-item North Carolina Teacher Evaluation Process (NCTEP) rubric. Responses were analyzed to understand the rater response process using the Many Facets Rasch Model (MFRM) for its evidence-based effectiveness in investigating reliability of education and psychological assessments (Bottoms, 2022).

Initial testing sought to establish the model-estimated proficiency levels for fabricated teacher profiles (Bottoms, 2022). Analysis of summative ratings on all 30 NCTEP rubric elements revealed a reasonable spread across profiles, indicating appropriate model-data fit and differing ability levels consistent with the expected field distribution (Developing [20% or two profiles], Proficient [60% or 6 profiles], and Accomplished [20% or 2 profiles]). Assessing variability in raters' strictness via a logit scale, MSE statistics indicated that some raters were consistently strict or lenient, while others exhibited more unpredictable rating patterns, highlighting irregularities in assessments.

Examination of group differences in rater severity based on race showed statistically significant variability in ratings. The overall bias effect was minimal, indicating a small overall group bias. However, individual rater patterns revealed that five raters exhibited bias, rating White teachers more leniently and teachers of color more strictly (Bottoms, 2022).

Assessment of the alignment of raters' grading trends on the fictitious profiles with their actual caseload, using the Summative Caseloads model, revealed an appropriate fit. This alignment was characterized by insignificant variability in teacher profiles and

illustrated by minimal change in logit location from one model to the other (Bottoms, 2022). Five raters demonstrated patterns of strictness or leniency in the field which were completely different from patterns observed in the study. Two raters demonstrated rating patterns in the field which were stricter, but not completely different, than those patterns observed in the study.

Potential consequences of identified bias in rater behavior impact the fairness and reliability of evaluations. Potential impartiality can lead to the denial of licensure due to demographic factors vice merit. Bottoms' (2022) results provide evidence for increased monitoring of raters employed by the EES Office, and the model aids in identifying areas where individual raters differ from the overall group patterns and prompt leadership to follow up with raters displaying statistics outside the expected or normal boundaries.

Other studies (Clemens et al., 2012; Bottoms, 2022; Dunst, 2018) and insights from after-action reports (Heuer, 2020) consistently advocate for enhanced rater training to address challenges in performance evaluations. However, Pufpaff, Clarke, and Jones (2015) challenge this consensus in their study, "The Effects of Rater Training on Inter-Rater Agreement." Employing a three-phase approach, the study assessed the consistency of ratings before and after directed training.

Ten full-time faculty members, varying in experience, volunteered for the study. They conducted rubric-based performance assessments on students in an undergraduate special education teacher training program (Pufpaff et al., 2015, p. 118). The raters evaluated student work both before and after receiving training on the rubric. The training materials were thoughtfully designed to accommodate faculty time constraints, prioritizing efficiency. These materials included an expanded rubric with additional details in each row, covering assignment requirements, knowledge, skills, dispositions, and/or performances aligned with professional standards (Pufpaff et al., 2015, p. 125). Written directions for candidates and definitions of rubric terms were also included (Pufpaff et al., 2015, p. 125). Accompanying the expanded rubrics were narrated PowerPoint presentations introducing the assignment, providing background information, and explaining each rubric row in detail (Pufpaff et al., 2015, p. 126).

Despite the comprehensive training, the analysis, which included side-by-side comparisons and focused on rater agreement, percentage agreement, and scores within one acceptable performance level of the true score (original rating by the course instructor), revealed limited agreement among raters. The rubric comprised 32 rows, resulting in 310 individual scores. Pre-training, approximately 43% of these scores concurred with the established true scores, and an additional 35% were within an acceptable range of the true score (Pufpaff et al., 2015). Only two rubric rows exhibited unanimous agreement, with all scores within one acceptable level of each other (Pufpaff et al., 2015). Post-training, the only change was a 3-percentage point decrease in the percentage of scores that fell within one acceptable level of the true score (Pufpaff et al., 2015).

Participants found the training highly useful, with 60% rating it as excellent and 40% as good (Pufpaff et al., 2015). After the training, participants expressed increased comfort in scoring assignments, with 40% rating the clarity of the materials as excellent and 60% as good (Pufpaff et al., 2015).

Pufpaff et al. (2015)'s study underscored the importance of rater training for achieving strong inter-rater reliability. However, despite these efforts, variability persisted in the assessment process, potentially impacting candidates' outcomes in the program. The study highlights the challenges of achieving consistent ratings in complex assessments, even with the implementation of training materials.

This study does have some limitations, particularly regarding transparency. It lacks details on the assignments assessed, does not provide copies of the grading rubric, and does not outline the statistical test or equation used for agreement calculation. The specific nature of the assignment and the details of the rubric can significantly impact grading subjectivity, offering more context to disparate ratings. In essence, the lack of transparency not only restricts the reproducibility of the study but also undermines its credibility. Pufpaff et al. (2015) argues for systematic training to improve assessment reliability. Moreover, their findings emphasize that training, despite being commonly perceived as a solution, may not always be as effective as believed in enhancing the consistency and dependability of performance evaluations or assessments. This suggests that the effectiveness of training

might hinge on the method and content employed, emphasizing their pivotal role in realizing tangible benefits.

B. MARINE CORPS PERFORMANCE EVALUATION CHALLENGES

This section covers studies that have been conducted into the PES which highlight the system's shortcomings. While some were intentionally examining RS reporting trends, some were conducted to investigate correlations between career paths and service member demographics/characteristics and subsequently identified potential bias in RS behavior. This focus is to be expected given the nature of a performance evaluation which is influenced not only by proficiency and performance but also personality and rapport. However, the unexplained scoring differences by demographic factors impacts reliability of the report and its reflection of the individual MRO's true quality and value to the organization.

1. Fitness Report Structure

The primary purpose of the FitRep is to aid in the decision-making of a promotion or special selection board, emphasizing an objective and accurate assessment of individual performance based on Marine Corps standards (Headquarters, United States Marine Corps, 2023). The PES manual underscores the importance of reporting facts and objective judgments, ensuring evaluations consider performance against established criteria, individual capacity, and professional character (Headquarters, United States Marine Corps, 2023, p. 1–2).

Jobst and Palmer (2005) introduce a critical perspective when they find evidence that contradicts the presumed objectiveness of the report and challenges the standardized nature of traits. They argue that reporting officials weigh each FitRep attribute differently based on their MOS. This revelation introduces a nuanced layer to the evaluation process, suggesting a need for a more tailored and weighted approach to competencies in the ongoing discourse on the effectiveness of evaluation tools.

Examining 33,858 OMPF records of officers (2nd Lieutenant through Colonel) from 1999 to 2004, Jobst and Palmer (2005) revealed significant variations in average

scores across MOSs and ranks, supporting the call for a weighted approach to performance evaluations. Mean scores of 14 competencies showed variances across MOSs, with, for instance, Financial Management Officers at the rank of 1st Lieutenant consistently scoring higher than other MOSs but displaying lower scores at the rank of Lieutenant Colonel (Jobst & Palmer, 2005). Proficiency, Initiative, and Communication also displayed higher averages in infantry, communications, and legal MOSs (Jobst & Palmer, 2005).

Jobst and Palmer (2005) employed an ordinary least squares regression to analyze FitRep competency averages, focusing on how estimated coefficients for MOS groups affect competency scores. Their findings suggested that MOS groups are significant factors in understanding competency scores, especially when considering the Marine's rank and the year of the report. Their results advocated for customized evaluations (Jobst & Palmer, 2005). They reasoned that the importance of a particular skill could be determined by assessing its statistically significant deviation in average FitRep score when compared to other MOSs, then the distinct skills, categorized as competencies or attributes, demanded by each MOS should be considered in the weighting of the associated FitRep competency (Jobst & Palmer, 2005).

Jobst and Palmer (2005) also conducted a survey (54.2 % response rate) focused on identifying competency importance in primary MOS duties. Respondents were primarily comprised of captains and majors, with overrepresentation of pilots and communication officers. Aside from rank and MOS, respondents were asked to rank each FitRep competency in terms of relevance to inherent duties and assigned duties of their billet. The survey results revealed consistent high ranking of Performance and Proficiency competencies across all MOSs, with consistent low rankings for Professional Military Education (PME) competencies. Additionally, significant variations were observed among MOSs in certain competencies, supporting the hypothesis that not all FitRep competencies hold equal importance across different MOSs.

Results of the Jobst and Palmer (2005) survey also highlighted the widespread belief, affirmed by 91 percent of officers, that skill sets should evolve as officers progress in pay grade, aligning with the observed increase in the importance of the Performance competency with higher pay grades. However, it is essential to note the lack of diversity in

MOS and grade of respondents limiting the generalizability of perceptions to the greater population of officers. This limitation is further aggravated by self-selection bias in the surveyed sample and the impact of non-randomized participation from fellow students, urging the need for a more comprehensive survey covering officers across the entire USMC, especially those actively serving in their primary MOS duties, to enhance the reliability of the findings.

Another aspect of the report that has raised concern is the relative value (RV). The PES manual (2023) describes the RV as a metric for promotion and special selection board members to weigh the merit of a single FitRep in relation to the RS profile. "A report's RV reflects how the fitness report average of an individual report compares to: (1) The RS's average of all fitness reports written by the RS on Marines of the same grade and (2) The highest fitness report average of any report written by the RS on a Marine of the same grade as the MRO." (Headquarters, United States Marine Corps, 2023, p. 8–6).

In his article "Stop Using Relative Values They Don't Work as Advertised," CNA research scientist and Marine Corps reservist Dr. Ryan Baker (2024) sheds light on the calculation and common misinterpretations of RVs. Using notional FitRep scores displayed as seen by board members, Baker's Table 3 illustrates the misleading nature of RVs. The fictional report had an RV of 80 at the time of processing (RV at Proc), suggesting it had never occupied the top position within the RS profile. Subsequently, the RS completed three additional reports (denoted by the boxes labeled "Reports" and "3 of 6"), leading to a cumulative RV escalation from 80 to 87.56 (Cum RV). This sequence highlights how RVs evolve over time, capturing the impact of subsequent reports on the cumulative assessment and, consequently, the challenges associated with interpreting these values in isolation.

Table 3. Sample reporting senior marks as presented on fictional MBS. Source: Baker (2024)

REPORTING SENIOR MARKINGS																
Reporting Senior		Per	Pro	Cou	Eff	Ini	Lea	Dev	Set	Ens	Co	PME	Dec	Jud	Eval
Promote	Reports	RPT Avg		RS Avg		RS High		RPT at High			RV at Proc			Cum RV	
		C	C	B	B	C	B	B	B	B	C	C	B	C	C
Yes	3 of 6	2.50		2.99		5.00		1			80			87.56	

Note: An extract from an MBS showing the attribute marks on a fitness report and how they compare to others in the RS profile. The RV is supposed to show the location of the report within the underlying RS profile, but it can't do so reliably (as Table 1 demonstrates).

Table 4 visually demonstrates the ambiguity in RV interpretation, portraying two distinct profiles generating the same RV observed in Table 3. Profile 2 places a report with an average of 2.50 at the RS profile's bottom, while in Profile 1, it is closer to the top. RVs cannot distinguish between these scenarios, posing challenges for selection boards in understanding a Marine's true position and performance.

Table 4. Two sample RS profile capable of producing the same RV. Source: Baker (2024)

	Profile 1	Profile 2
RS Profile		
Report 1	4.00	2.57
Report 2	5.00	2.64
Report 3	**2.50**	**2.50**
Report 4	2.08	2.62
Report 5	2.00	2.62
Report 6	2.36	5.00
Master Brief Sheet		
Reports	6	6
RPT Avg	2.50	2.50
RS Avg	2.99	2.99
RS High	5.00	5.00
RPT at High	1	1
RV at Proc	80.00	80.00
Cum RV	87.56	87.56

Note: Two RS profiles, either of which could have produced the numbers in Figure 1. In Profile 1, the report highlighted in pink is above the median. In Profile 2, the same report is last.

Drawing an analogy to Anscombe's Quartet in statistics (four datasets that appear uniform in basic descriptive statistics but possess distinct qualities) (Siegrist, 2022), Baker (2024) highlights how FitReps with the same report average can create the illusion of uniformity on the MBS. Table 5 highlights the loss of information in summarizing data. There are several FitReps, spread across five RS profiles, which have identical cumulative RVs on the MBS. However, when the reports are shown in the RS profile, distinctive patterns emerge, revealing significant differences in their underlying distributions. Baker (2024) uses this table to emphasize the challenge of inferring the report's location within the profile from MBS information alone, analogous to the diverse visual patterns observed in Anscombe's Quartet.

Table 5. Illustration of equivalent FRAs appearing in distinct rs profiles. Source: Baker (2024)

	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5
RS Profile					
Report 1	5.00	5.00	5.00	5.00	5.00
Report 2	3.38	3.54	4.00	4.77	**3.00**
Report 3	3.31	3.46	3.38	**3.00**	**3.00**
Report 4	3.23	3.31	**3.00**	2.77	**3.00**
Report 5	3.08	**3.00**	2.92	2.77	
Report 6	**3.00**	2.69	2.69	2.69	
Master Brief Sheet					
Reports	6	6	6	6	4
RPT Avg	3.00	3.00	3.00	3.00	3.00
RS Avg	3.50	3.50	3.50	3.50	3.50
RS High	5.00	5.00	5.00	5.00	5.00
RPT at High	1	1	1	1	1
RV at Proc	-	-	-	-	-
Cum RV	86.67	86.67	86.67	86.67	86.67

Note: This table shows the disconnect between the RS profile and MBS. The top portion shows five alternative RS profiles. The bottom portion shows how the highlighted reports in each profile are represented on the MBS. RSs can see the top portion; selection boards can see the bottom portion. Neither can see what the other sees.

Baker (2024) identifies two critical factors contributing to RV misinterpretation. First, the asymmetry of the RV scale challenges the belief in its symmetry (Gupta, 2022), allowing even the lowest-ranked report to achieve an RV higher than 80 (Baker, 2024).

Second, replacing the highest report in the profile with a higher one extends the RV scale in both directions, counterintuitively elevating RVs of reports at the profile's bottom (Baker, 2024).

Furthermore, Baker (2024) highlights the challenge faced by selection boards in discerning whether an RV change is due to reports added above or below it in the profile, introducing ambiguity. The order of report processing and the uneven distribution of reports in RV "thirds" add complexity to interpretation.

Baker's (2024) critique extends to three critical assumptions made by selection boards regarding RVs: the assumption of the profile average being centered, meaningful performance differences, and a reliable proxy for the RS's marking philosophy. He argues that RV accuracy depends on the weak law of large numbers, emphasizing the need for a large sample size for meaningful patterns (Baker, 2024).

Contradicting the Marine Corps' perspective on RVs, Baker (2024) asserts that RVs significantly influence selection probability and challenge the notion that downplaying their importance suffices for fair evaluations. To address these issues, he proposed replacing RVs with graphical representations of RS profiles on the MBS, aiming to provide a clearer understanding of profile quirks compared to summary statistics.

These studies reveal complexities in performance evaluations. Jobst and Palmer's (2005) insights call for a nuanced, MOS-specific approach to competencies, challenging the one-size-fits-all model. Baker's (2024) research critiques the RV system, emphasizing its misleading nature and proposing alternative visual representations for a clearer understanding of RS profiles. These studies collectively emphasize the need for ongoing refinement and customization in the evaluation process to ensure true and accurate assessments of quality performance.

2. Reporting Senior Markings

Fair and impartial assessments are additional indicators of effective performance evaluation systems. Yet, previous research into the Marine Corps PES has found evidence of potential bias and effects of poor training on FitRep scores (Clemens et al., 2012). Using

a combined approach inclusive of a comprehensive analysis of all officer FitReps from January 1999 to August 2011, review of the FitRep training curriculum, and stakeholder interviews, Clemens et al. (2012) were able to identify the system's strengths, it's challenges and propose recommendations for improvement.

Findings suggested several aspects of the system were performing as intended. Grade inflation was dismissed, finding that FRAs experienced a slight increase until FY03, followed by a subsequent decline (Clemens et al., 2012). Promotion of the best-qualified officers was affirmed via agreement between subject matter experts.

Clemens et al. (2012) identified multiple areas of the PES warranting concern. Findings revealed a disparity between the intended and observed distributions of RO marks. "RO marks are intended to have a distribution referred to as them 'Christmas tree,' with few marks at the top in order to help boards identify exceptionally qualified Marines." (Clemens et al., 2012, p. 2). However, Clemens et al.'s (2012) analysis revealed the actual distribution deviated notably, as officers in higher ranks received higher RO marks on average (Clemens et al., 2012). Jobst and Palmer's (2005) findings suggest that this deviation could be attributed to the escalating responsibility and accountability accompanying higher ranks. They propose that increased scores reflect the assumption that skill sets evolve and improve with growing leadership responsibilities.

Analysis revealed a correlation between FitRep scores, entry paths and academic achievements (Clemens et al., 2012). Officers commissioned through E-to-O programs or with higher college GPAs consistently earned higher FitRep scores (Clemens et al., 2012). Furthermore, the influence of school quality on performance evaluations is less significant compared to the impact of college GPA (Clemens et al., 2012).

Analysis of FitRep scores among different racial groups revealed that white RSs assigned slightly lower FRAs to black MROs and vice versa (Clemens et al., 2012). Furthermore, Clemens et al. (2012) identified racial disparities in promotion recommendations. Non-White officers, with comparable RVs, received less favorable endorsements than White officers (Clemens et al., 2012).

Analysis also revealed the potential for bias towards officers in specific occupational fields, such as aviation, who consistently received lower FitRep scores. Clemens et al. (2012) suggested further investigation to develop a comprehensive understanding of the factors contributing to the disparities, to uncover any underlying biases and discern potential systemic issues contributing to patterns of inconsistencies and variance.

Dunst (2018) found similar observations in RS scoring trends in his thesis on the evolution of Marine Corps FitReps. Statistical analysis of 118,765 FitReps, spanning from FYs 2010 to 2017, uncovered several correlations and patterns of education and demographics on FitRep outcomes (Dunst, 2018).

To estimate the probability of an MRO being rated in the top third on a FitRep based on RS and MRO characteristics, Dunst (2018) used multivariate logistics regressions. Performance-based factors were found to play a crucial role in determining FitRep outcomes, with physical fitness test scores and combat experience serving as influential components (Dunst, 2018). However, education consistently emerged as a primary and noteworthy factor shaping the MRO assessment, especially within science, technology, engineering, and mathematics (STEM) fields. Education stood out as the most influential predictor of MROs receiving top-third FitRep outcomes with significant varied effects across occupational fields. At the intersection of education and gender, female officers with degrees in STEM fields are identified as less likely to receive top-third FitRep outcomes (Dunst, 2018).

A significant racial correlation is identified, revealing that White MROs tend to receive more favorable ratings, particularly when assessed by White RSs. Conversely, non-White RSs exhibit a tendency to rate non-White MROs relatively lower. However, the study highlights that when factors like education level and combat experience are taken into account, the significance of racial correlations becomes less pronounced.

The analysis further explored the learning curve exhibited by RSs over time. The variation in RS FitRep outcomes notably narrows when accounting for performance-based factors. This narrowing suggests a learning process for RSs, indicating an increased

proficiency in evaluating the performance of MROs over the course of their roles. The recognition of this learning curve emphasizes the dynamic nature of the evaluation process and implies that RSs become more adept at discerning performance nuances with accumulated experience.

Two important themes emerge from this research. First is how an officer's background, such as race, gender, or education correlates to their evaluation. Both studies found racial disparities in recommendations and marks. Dunst (2018) found disparities in gender and education while CNA found further disparities in commissioning source (Clemens et al., 2012). Second is the call for more comprehensive training as the solution to score variations.

The limitations of the Dunst and CNA studies are evident in several aspects. First, both studies heavily rely on quantitative methods, such as logistic regression models, which may provide statistical correlations but lack the richness of qualitative insights. Qualitive insights from RSs on behavior indicative of attribute and rating, or other personality traits which contribute to performance evaluation but are not included on the FitRep could reveal subjective elements and decision-making factors not captured by quantitative measures. Additionally, both studies have a limited exploration of non-demographic factors such as the nature of assignments, specific achievements, or interpersonal skills, which could play a significant role in FitRep outcomes. Furthermore, the studies fall short in establishing causation or the nature of the interrelationships. While correlations between factors like race, education, and gender and FitRep outcomes are identified, the reasons behind these correlations are not deeply explored. Exploring qualitative information and non-demographic factors would help provide context to the statistical outcomes.

Addressing another aspect of the PES which can be a source of conflict or disparity, Rigaut (2017) conducted a text analysis of Marine Corps FitRep section I and K comments. In a distinctive approach, Rigaut (2017) analyzed a dataset comprised of over 71,000 FitReps from officer cohorts in 1996, 1997, 2006, and 2007 using text statistics and machine learning algorithms. Findings revealed that well-written reports using simple words in longer sentences, with an emphasis on future command opportunities, indicated the best performers (Rigaut, 2017). However, challenges in RV and comparative

assessments (CA), along with inconsistent evaluations, suggest the need for standardized language in fitness reports (Rigaut, 2017).

The study aimed to analyze the informational value of text fields in fitness reports, focusing on the relationship between textual information and officer performance tier classifications. Using text mining, readability assessment, and supervised machine learning models to predict officer performance tier classifications, Rigaut (2017) was able to quantify the relationship between markings and comments, emphasizing the quality of the response variable.

Analysis of the RV and CA align with findings from Clemens et al. (2012) demonstrating a deviation from intended scoring distributions. Rigaut's (2017) findings revealed the RV and CA having a narrow distribution range. The RV distribution concentrated in the upper half, with higher ranks exhibiting higher values (Rigaut, 2017). The CA distribution, intended to resemble a "Christmas Tree," also deviated from the expected pattern, with lieutenant colonels receiving higher marks than second lieutenants (Rigaut, 2017).

Highlighting concerns raised by Clemens et al. (2012) and Baker (2024), these trends impact the Marine Corps' ability to distinguish talent within the officer ranks. Rigaut's (2017) findings demonstrate an additional complexity faced by the board in identifying clear separations between performance tiers, leading to potential misclassifications of fitness reports. The study also identifies discrepancies between the intended and actual distributions of these assessment metrics, raising concerns about their effectiveness in accurately reflecting officer performance. Rigaut (2017) suggests adoption of standardized language to ensure consistency in Section I and K comments by "enhancing the word-picture guidance to separate talented Marines and promote conformity in issuing quantitative assessments of performance." (p. xxi).

Analysis of the concurrence between RS and RO in assigning MROs to performance tiers highlighted another significant disparity. In a sample of 71,212 FitReps, less than 0.1% exhibited non-concurrence (105 reports) meaning the RO agreed with the RS's assessment of the MRO's performance and subsequent PARS ratings (Rigaut, 2017).

However, a substantial 49% disparity was observed in assigned tiers (Rigaut, 2017). Differences in tier assignments were calculated, revealing a 43.1% disagreement by one tier and 6.1% by two tiers, with ROs marking higher than RSs in 27.1% and RSs higher than ROs in 22.1% of cases (Rigaut, 2017).

Rigaut (2017) suggests that reasons for the disparity include RS marking based on MRO's performance during the reporting period, while RO considers comparisons to all Marines of the same grade known professionally to the RO.

In the exploration of RS markings Clemens et al. (2012) and Dunst (2018), revealed significant patterns and disparities. Furthermore, Rigaut (2017) identified a significant disparity in tier assignments between RS and RO adding another layer of complexity to the evaluation process. Together, these studies emphasize the multifaceted nature of RS markings, the impact of various factors on FitRep outcomes, and the need for continuous improvement in the evaluation system.

C. SUMMARY

The literature review explores effective performance evaluation measures, covering Chachula (1992) and Small's (2020) studies on common characteristics and best practices in successful companies. It introduces the relative versus absolute measures paradigm, highlighting challenges associated with relative assessments. Balakrishnan, Lin, and Sivaramakrishnan (2016)'s study explores the efficacy of absolute and relative systems, providing insights into task assignment and system preferences. Content validity and accuracy are addressed through Luke's (2022) examination of the U.S. Navy's trait value statements, while Bottoms' study (2022) emphasizes the need for reliability and uncovers variations in evaluators' strictness. Jobst and Palmer's (2005) focus on FitRep structure underscores the importance of a weighted approach, and the review addresses bias in performance evaluations, drawing attention to RS markings and disparities in FitRep scores. The complex dynamics influencing Marine Corps performance evaluations are explored through studies by Clemens (2012), Dunst (2018), years needed and others, offering a comprehensive overview of performance evaluation measures in organizational and military contexts.

D. LITERATURE CONNECTION TO USMC PERFORMANCE EVALUATIONS

These studies collectively contribute to an understanding of effective performance evaluation measures, the challenges posed by relative and absolute assessments, the importance of content validity and accuracy, and the complexities and potential biases in Marine Corps performance evaluations.

With the current system in place for over two decades and a substantial body of research highlighting significant issues in evaluation rating patterns, the next logical step is an in-depth review of the system to better understand and validate the findings of previous research. The literature offers diverse frameworks for the Marine Corps to assess the PES, encompassing empirical reviews to align the FitRep with organizational doctrine and incorporate best practices from sister services and successful businesses in diverse industries. Additionally, it demonstrates the value of conducting validity and reliability testing to ensure that score variations are not predominantly a result of a poorly designed system, introducing inconsistencies and unreliable assessments of performance and quality.

Studies on the Marine Corps PES highlight the influence of observable factors on RS scoring. The focus of recommendations is on making changes to training, the system itself, and the measurement tools used in the evaluation process. Exploring system modifications such as the implementation of a weighted FitRep to better illustrate the value of different skills, does hold significance. However, this focus should be subsequent to a comprehensive assessment of the entire system, including statistical analysis, before adopting any changes. This approach aims to identify the root cause of the issues rather than reacting to correlated factors.

By investigating the degree of agreement or consistency among RSs when assessing the same Marine's performance, this thesis provides empirical evidence on the reliability of the FitRep. This is an essential contribution as inter-rater reliability is a fundamental psychometric property that ensures the consistency and fairness of evaluations.

IV. METHODOLOGY

A. PURPOSE AND RESEARCH QUESTION

Our research aims to explore the reliability of the USMC FitRep by analyzing RS inter-rater reliability. While acknowledging that inter-rater reliability is not the sole determinant of product quality, it stands as a crucial criterion in the subjective evaluation of product quality, as highlighted by Nichols et al. (2010). This chapter explains the methodology used to conduct the analysis.

B. POPULATION AND SAMPLE

The population includes active-duty United States Marines grades O1 through O7. This effectively encompasses service members who inherently serve as RSs under the PES. It is also the population of service members who receive a FitRep and are subject to a comprehensive boarding process. Enlisted service members ranked E-4 through E-9 also receive FitReps but are excluded from the population as they cannot serve as an RS (except in special circumstances) and given their distinctive promotion board process which considers the merits of service members by MOS category.

The target sample for this study was limited to officers in grade O1 to O5. General Officers in grades O6 and O7 form a minority within the target population and typically serve in Command positions. Consequently, they are more likely to serve as an RO and thereby were excluded from the targeted sample.

The survey opened on January 31, 2024, and closed on February 25, 2024. Participation was solicited via email to officers fitting the sample demographic. See Appendix D for copy of email invitation. The invitation was extended primarily to officers enrolled at the Naval Postgraduate School (NPS), with the Marine Corps student body totaling 273 students.

C. DATA COLLECTION

To rigorously test reliability in a simulated scenario, a randomized control trial would have been ideal. This trial would involve multiple groups of officers, at least three,

serving as RSs. Participants would be randomly assigned to these groups, each tasked with evaluating a minimum of three distinct MROs. These MROs would be stratified based on their true RSs profile, with one in the bottom third, another in the middle, and one in the top third. Additionally, assignments of MROs to groups would also be randomized. Participants would also conduct a second evaluation with randomized assignment of MROs to assess consistency of ratings over time.

This design would facilitate the assessment of both within-group reliability, examining consistency of ratings within each group, and between-group reliability, evaluating consistency of ratings between different groups. Furthermore, random assignment of officers to groups would help mitigate potential biases or preferences.

FitRep individual attribute ratings, report averages and rating justifications would be compared to that of the true RS. By comparing assessments made by the experimental groups against the true ratings provided by actual RSs, researchers can directly measure the accuracy and reliability of the evaluations. This comparison serves as a benchmark for evaluating the effectiveness of the simulated assessments.

However, conducting a randomized control trial with multiple groups of officers would have been logistically challenging in this situation due to constraints such as limited timeframe, resources, and access to personnel. Ensuring blinding and preventing biases in such a trial would have been immensely difficult within the military context, as officers may have pre-existing knowledge or biases about the individuals being evaluated, the billet, unit or the evaluation process itself.

Alternatively, another method for studying test-retest and inter-rater reliability could have involved using MBS data from MMRP. This approach would allow for analysis of multiple FitRep evaluations from the same RS over time to examine test-retest reliability. Similarly, it would enable analysis of FitRep evaluations from multiple RSs on multiple MROs with similar characteristics (i.e., rank, billet and unit type) to test inter-rater reliability. Both methods facilitate the analysis of real-world data and enable the necessary direct observation to assess performance more accurately. However, challenges arise from the lack of information on PARS ratings adjusted based on comparisons with

the RS profile, and the inability to gather qualitative insights from RSs regarding their rationale or decision-making process behind the ratings.

Given these limitations and constraints, we opted for a one-time survey and scenario approach, simulating a first-time FitRep evaluation with minimal observation. Our data collection involved two primary components: a customized survey and fictional FitRep evaluations.

1. Survey

Utilizing an online survey for this research offered numerous advantages. First, it eliminated a cost for distribution and manhours/training for interviewers: "Online surveys are characterized by speed, reach, ease, flexibility, and automation" (Ball, 2019). Qualtrics allowed for ease in editing and construction in addition to multiple format export of the results. Respondents were able to access the survey from any geographical location with internet access and complete it at their own convenience within the given timeframe. The advantage for us was efficiency and the assumption that respondents are more likely to provide honest and open responses when participating in an online survey bolstering the reliability of the collected data.

There are several drawbacks associated with the use of online surveys for research purposes. Among these are incomplete responses, which significantly impacted our sample size and the number of usable observations, thereby affecting the generalizability of our results. Additionally, limitations include the inability to conduct follow-up questions with an interviewer, the potential for sample bias when utilizing virtual platforms for distribution, challenges in detecting survey fraud, and the ambiguous nature of anonymity, which can both encourage honest responses and facilitate fraudulent behavior (Ball, 2019).

Although the absence of follow-up questions may limit the depth of understanding gained from respondents, we attempted to address this limitation by incorporating open-ended questions to justify the PARS assessment ratings. Additionally, given the constraints of sampling and the inability to employ optimal random sampling techniques such as participant selection or recruitment (Ball, 2019), we opted for an alternative approach by inviting a predefined population to participate through email outreach.

While acknowledging the limitations inherent in surveys, we deemed a survey the most appropriate method for this research. An alternative approach, similar to Larger's 2017 study, involved quantitative analysis testing reliability by examining MBSs to investigate the consistency of evaluation scores among multiple MROs with similar billet descriptions. However, this approach was less preferred due to challenges such as unidentified or unannotated accomplishments potentially influencing scoring variations, the lack of access to RS profiles for determining marking philosophy, and other undisclosed factors that might contribute to rating disparities.

To investigate our research question, we developed a tailored electronic survey using Qualtrics. See Appendix C for the survey instrument. The survey consisted of 13 questions separated into three sections:

1. Section one included informed consent providing a legal notice and advising potential participants of their right to privacy.

2. Section 2 collected background information from participants including rank, primary MOS, completion of evaluation training and number of Marines reported on as an RS.

3. Section three incorporated a practical scenario where participants stepped into the role of an Executive Officer and were tasked with completing an annual FitRep for a subordinate staff member. Respondents accessed a completed MROW and used it to evaluate the MRO's accomplishments. Lastly, participants were asked to provide an overall assessment rating.

a. Survey Scenario

In recognition that respondents may not possess a detailed understanding of the duties associated with the simulated MRO position and did not have direct observation, we crafted the survey scenario to align with a billet that mirrors the potential circumstances of the respondents. Command billets, to include Executive Officer positions, are typically held by Marine Officers in the O2-O5 grade range, depending on the command level. In this capacity, they serve as an RS for multiple MROs from diverse MOSs, which they may

not have had prior interaction with and are unaware of the full scope of associated duties. In addition, the Executive Officer works in the HQ building and may be geographically separated from some of his or MROs, preventing direct observation. The choice for respondents to assume this role aims to inject realism into the scenario, mimicking potential challenges related to limited observation and lack of prior knowledge of MRO billet responsibilities.

b. Attribute Selection

We designed section three of the survey to mirror sections D through G of the current NAVMC 10835 (Rev. 7–11) FitRep, focusing on eight attributes out of the 14. The surveyed attributes included Performance, Courage, Initiative, Setting the Example, Communication, PME, Decision-Making Ability, and Judgment. This intentional reduction was designed to streamline the survey completion process, allowing participants to offer more thoughtful and precise responses while minimizing the risk of survey fatigue. Notably, the chosen attributes were ones that can be reasonably inferred from the MROW without a strict requirement for direct observation.

Attributes such as Proficiency, Effectiveness Under Stress, Leading Subordinates, Developing Subordinates, and Evaluations were purposefully omitted. The exclusion of Proficiency, Effectiveness Under Stress, and Leading Subordinates was motivated by the challenges associated with measuring these attributes without direct observation. Additionally, Evaluations, which gauges the commitment of the RS to accurate, unbiased, and punctual evaluations, was omitted due to its distinctive nature.

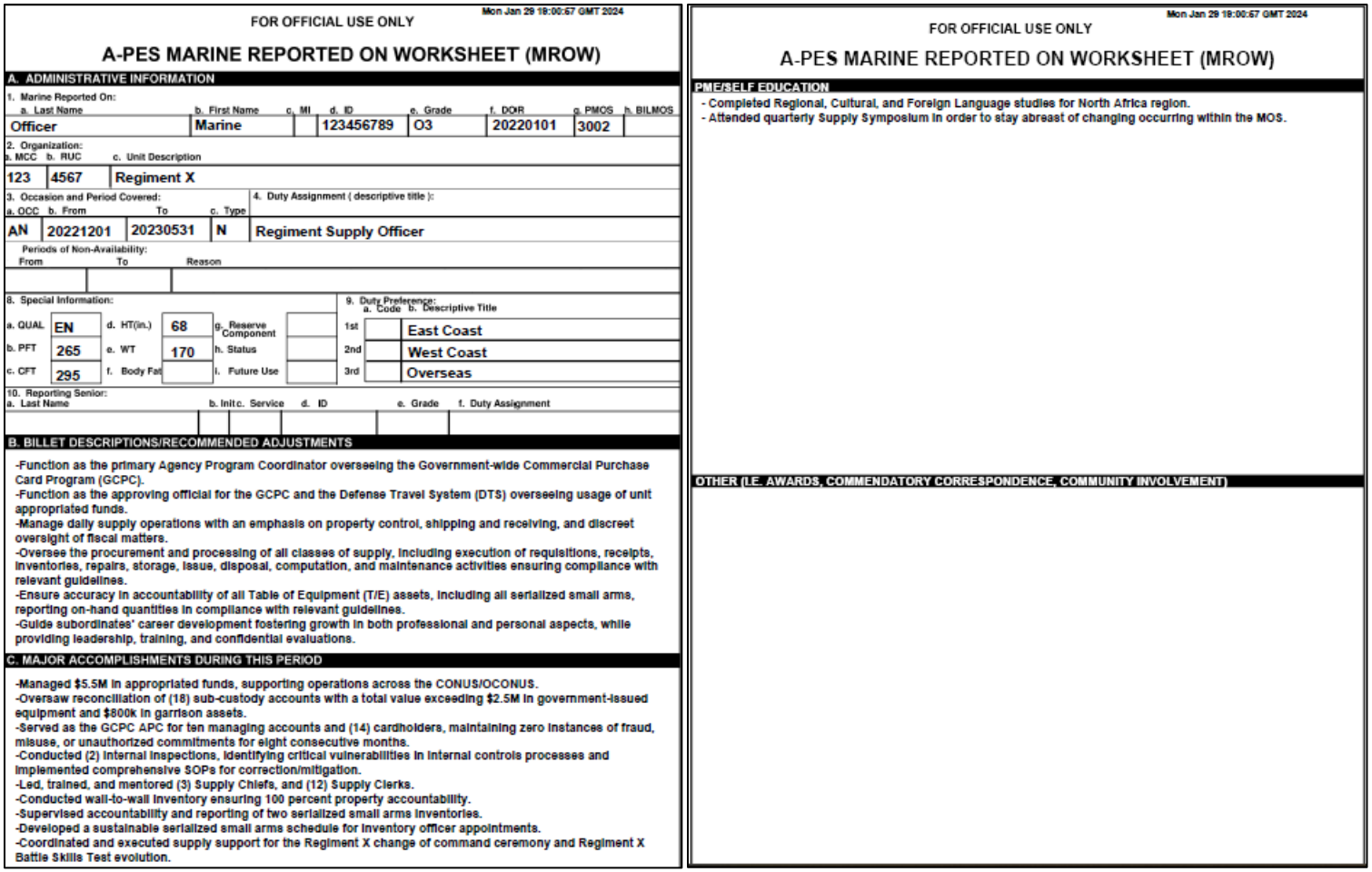
2. MROW

The MROW utilized in the survey originated from archived FitReps submitted by a company-grade officer. The officer, after signing a consent form, submitted two completed FitReps via email for the purpose of generating the fictional MROW for the survey. The officer received detailed information from us about the research nature through email communication and explicitly provided consent for the utilization of the reports to create a fictitious evaluation.

To ensure anonymity and confidentiality, Section A information underwent fictionalization. This effort involved using a generic name to avoid gender indication, entering Electronic Data Interchange Personal Identifiers (EDIPI) and unit identifiers as sequential numbers to eliminate identification, and describing annual training requirements as "reasonable" first-class scores. Height and weight were entered within an ambiguous range to prevent gender inference. The unit name was labeled as "Regiment X" to provide context to command size and structure without specifying a particular unit. The Date of Rank (DOR) positioned the MRO at one year and five months Time-in-Grade (TIG) when the report was written, allowing for a six-month observation period for the Executive Officer, considering the potential impact of limited observation on respondents' assessments.

We also manipulated sections B and C of the two FitReps. Selection of billet responsibilities and accomplishments was based on their alignment with the overarching expectations associated with the MOS and particular billet. Our approach aimed to maintain a fair and unbiased assessment by avoiding the inclusion of highly specialized or exceptional requirements that might introduce potential recognition or bias in the evaluation process ensuring an impartial analysis.

Additionally, we used ChatGPT to modify the wording within the FitRep, crafting achievements that were both fictionalized and realistic. See Figure 1 for final MROW used in survey.



Mon Jan 29 19:00:57 GMT 2024

FOR OFFICIAL USE ONLY

A-PES MARINE REPORTED ON WORKSHEET (MROW)

A. ADMINISTRATIVE INFORMATION

1. Marine Reported On:

a. Last Name	b. First Name	c. MI	d. ID	e. Grade	f. DOR	g. PMOS	h. BILMOS
Officer	Marine		123456789	O3	20220101	3002	

2. Organization:

a. MCC	b. RUC	c. Unit Description
123	4567	Regiment X

3. Occasion and Period Covered:

a. OCC	b. From	To	c. Type	4. Duty Assignment (descriptive title):
AN	20221201	20230531	N	Regiment Supply Officer

Periods of Non-Availability:

From	To	Reason

8. Special Information:

a. QUAL	EN	d. HT(in.)	68	g. Reserve Component	
b. PFT	265	e. WT	170	h. Status	
c. CFT	295	f. Body Fat		i. Future Use	

9. Duty Preference:

	a. Code	b. Descriptive Title
1st		East Coast
2nd		West Coast
3rd		Overseas

10. Reporting Senior:

a. Last Name	b. Init	c. Service	d. ID	e. Grade	f. Duty Assignment

B. BILLET DESCRIPTIONS/RECOMMENDED ADJUSTMENTS

-Function as the primary Agency Program Coordinator overseeing the Government-wide Commercial Purchase Card Program (GCPC).
-Function as the approving official for the GCPC and the Defense Travel System (DTS) overseeing usage of unit appropriated funds.
-Manage daily supply operations with an emphasis on property control, shipping and receiving, and discreet oversight of fiscal matters.
-Oversee the procurement and processing of all classes of supply, including execution of requisitions, receipts, inventories, repairs, storage, issue, disposal, computation, and maintenance activities ensuring compliance with relevant guidelines.
-Ensure accuracy in accountability of all Table of Equipment (T/E) assets, including all serialized small arms, reporting on-hand quantities in compliance with relevant guidelines.
-Guide subordinates' career development fostering growth in both professional and personal aspects, while providing leadership, training, and confidential evaluations.

C. MAJOR ACCOMPLISHMENTS DURING THIS PERIOD

-Managed $5.5M in appropriated funds, supporting operations across the CONUS/OCONUS.
-Oversaw reconciliation of (18) sub-custody accounts with a total value exceeding $2.5M in government-issued equipment and $800k in garrison assets.
-Served as the GCPC APC for ten managing accounts and (14) cardholders, maintaining zero instances of fraud, misuse, or unauthorized commitments for eight consecutive months.
-Conducted (2) internal inspections, identifying critical vulnerabilities in internal controls processes and implemented comprehensive SOPs for correction/mitigation.
-Led, trained, and mentored (3) Supply Chiefs, and (12) Supply Clerks.
-Conducted wall-to-wall inventory ensuring 100 percent property accountability.
-Supervised accountability and reporting of two serialized small arms inventories.
-Developed a sustainable serialized small arms schedule for inventory officer appointments.
-Coordinated and executed supply support for the Regiment X change of command ceremony and Regiment X Battle Skills Test evolution.

Mon Jan 29 19:00:57 GMT 2024

FOR OFFICIAL USE ONLY

A-PES MARINE REPORTED ON WORKSHEET (MROW)

PME/SELF EDUCATION

- Completed Regional, Cultural, and Foreign Language studies for North Africa region.
- Attended quarterly Supply Symposium in order to stay abreast of changing occurring within the MOS.

OTHER (I.E. AWARDS, COMMENDATORY CORRESPONDENCE, COMMUNITY INVOLVEMENT)

Figure 1. Completed MROW presented to survey respondents for evaluation

D. DATA ANALYSIS

We conducted our analysis using a combination of R and Excel. The initial focus was on understanding respondent demographics through various methods followed by statistical reliability and text analysis.

1. Respondent Demographics

We employed various methods to derive descriptive statistics that revealed key patterns and trends among respondents. Initially, we determined the rank distribution by aggregating and categorizing the responses related to participants' military ranks. This approach allowed for a comprehensive overview of the categorized composition of the sample. Similar formulation was conducted for MOS category (survey question 3), FitRep training level (survey question 4) and for FitRep evaluation experience (survey question 5).

2. Quantitative Analysis of PARS Ratings and Averages

Subsequently, to gauge the overall sentiment of the MRO's performance, we converted the PARS ratings into numerical values for quantitative analysis. The conversion facilitated the calculation of a report average for each respondent, providing a consolidated measure of their evaluations. Additionally, we treated the overall assessment markings (question 15), as factors to explore possible relationships. Visual representations, including column, line and cluster charts, were generated to visually convey the overall distribution and distribution by specified respondent demographics.

3. Fleiss' Kappa

Reliability is integral to performance assessments, determining the trustworthiness of scores in conveying meaningful information (National Research Council, 1991). It gauges the consistency and dependability of assessment scores, ensuring stability and accuracy (National Research Council, 1991). Consistency implies stable scores with minimal variability, and dependability ensures the assessment instrument provides reliable information about an individual's performance.

The optimal approach to assessing reliability involves subjecting individuals to repeated testing with equivalent measures, aiming for nearly identical scores (National Research Council, 1991). A resulting small standard deviation relative to the population or potential score range indicates reliability (National Research Council, 1991). In research, calculating test reliability involves having two or more raters perform at least one repetition of an equivalent test. In cases where two raters assess examinee performance, the consistency of raters can be evaluated through intercorrelating their scorings, yielding an index known as interrater reliability (National Research Council, 1991).

In this study we employed the Fleiss' Kappa statistic as a robust measure of reliability. This statistical tool helped us to unravel patterns of consistency among evaluators, serving as a means to answer the primary research question centered on agreement among raters, thereby indicating the dependability and reliability of the scores assigned.

The Fleiss' Kappa statistic has strength in its ability to assess inter-rater consensus between more than two raters for categorical measurements (Nichols et al., 2010). Reliability is assessed by the level agreement as determined by the Kappa coefficient. See Figure 2 for coefficient definition.

Fleiss Kappa	Interpretation
<0.00	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect

Figure 2. Fleiss' Kappa agreement levels and definitions. Source: Sreedhara (2015)

The Excel data was read into R to assess inter-rater reliability. The subsequent steps focused on data transformation wherein the dataset was transposed to arrange raters or respondents as rows and PARS ratings as columns. This transposed table was then

converted into a numeric version by replacing letter grades with corresponding numeric values.

We computed Fleiss' Kappa using both the original transposed table and the numeric version. This statistical measure provided insights into the inter-rater reliability using a categorical measurement, offering a quantifiable assessment of agreement among respondents for the evaluated attributes in the survey responses.

4. Intraclass Correlation Coefficient

The ICC quantifies the consistency of rankings or measurements over multiple observations, indicating how well subjects maintain their relative positions across repeated measurements in addition to systematic differences between raters (Liljequist et al., 2019). For this study, we used the ICC to accommodate ordinal classifications derived from an analysis of variance (ANOVA) model. The ICC is calculated using the mean squares derived from the ANOVA model, including mean squared sum of squares between subjects (MSB), mean squared sum of squares between raters (MSJ), and mean squared error (MSE) (Mitani et al., 2017). Of note, the effectiveness of the ICC may be constrained by individual variations within subjects, random errors in measurements, and systematic biases inherent in the measurement method (Liljequist et al., 2019).

The ICC yields a score ranging from 0 to 1, with higher scores signifying increased consensus (Liljequist et al., 2019). We employed the two-way random-effects model. The model is commonly used in reliability studies to distinguish between variability within and between raters and aims to generalize results to raters who share similar characteristics with those included in the study (Koo & Li, 2016)

1. Within Raters (Intra-Observer Variability) measures consistency or the extent to which a single rater's assessment may differ across repeated evaluations.

2. Between Raters (Inter-Observer Variability) measures agreement or the degree to which various raters provide different evaluations for an identical set of items.

The ICC was calculated in R. The same two-way table, classifying the raters and the items as the two main factors, generated for the Fleiss' Kappa test was used to calculate the ICC. To assess the proportion of total variability attributable to rater differences, the ICC was calculated by partitioning the total variance observed in the data into two components: the variability within raters (consistency) and the variability between raters (agreement).

5. Text Analysis

The process of text analysis was executed to reveal insights beyond the scope of traditional quantitative analysis. Recognizing the inherent complexity of qualitative data, our approach allowed us to uncover patterns, recurring themes, and underlying connections within the text, providing a depth of understanding that goes beyond the confines of numerical representation.

Initially, data was extracted from Qualtrics in an Excel format. Subsequently, the Excel sheet was filtered to spotlight columns containing both PARS ratings (survey question 6) and accompanying PARS justification comments (survey questions (7–14). To ensure a structured approach, the data was organized by attribute and further categorized by PARS rating. The review process included carefully examining each comment to identify recurring words, phrases, and accomplishments used to justify specific ratings. Additionally, the comments were analyzed to identify performance measures and personality traits that respondents deemed essential or beneficial for a more accurate assessment of each attribute. This systematic and comprehensive methodology was employed to derive nuanced insights from the survey responses.

E. SUMMARY

This study aimed to investigate the reliability of the USMC FitRep by analyzing RS inter-rater reliability. Acknowledging the critical role of inter-rater reliability in evaluating product quality, our methodology was designed to analyze the consistency and dependability of FitRep assessments. We targeted active-duty United States Marines in grades O1 through O5 as our population of interest, excluding enlisted service members and higher-ranking officers due to their distinct roles within the evaluation process.

Data collection involved the distribution of a tailored electronic survey to officers fitting the sample demographic, primarily targeting those enrolled at NPS. The survey incorporated a practical scenario where participants evaluated a fictional FitRep for a subordinate staff member. Data analysis encompassed quantitative analysis of PARS ratings and averages, Fleiss' Kappa for inter-rater reliability, ICC for consistency, and text analysis to uncover qualitative insights. The next chapter presents the findings obtained from our data analysis.

V. ANALYSIS

In this chapter, we present an overview of the respondent statistics derived from the survey, providing context and insight into the participant demographics. The examination of respondent demographics and training backgrounds lay the groundwork for understanding their perspectives and insights into the subsequent analysis of inter-rater reliability and text analysis.

A. RESPONDENT DEMOGRAPHICS

Responses to the survey were received from 76 officers, constituting a 28% response rate. The total population of Marine Corps Officers within the NPS student body at the time of the survey stood at 273. See Figure 3 for rank dispersion.

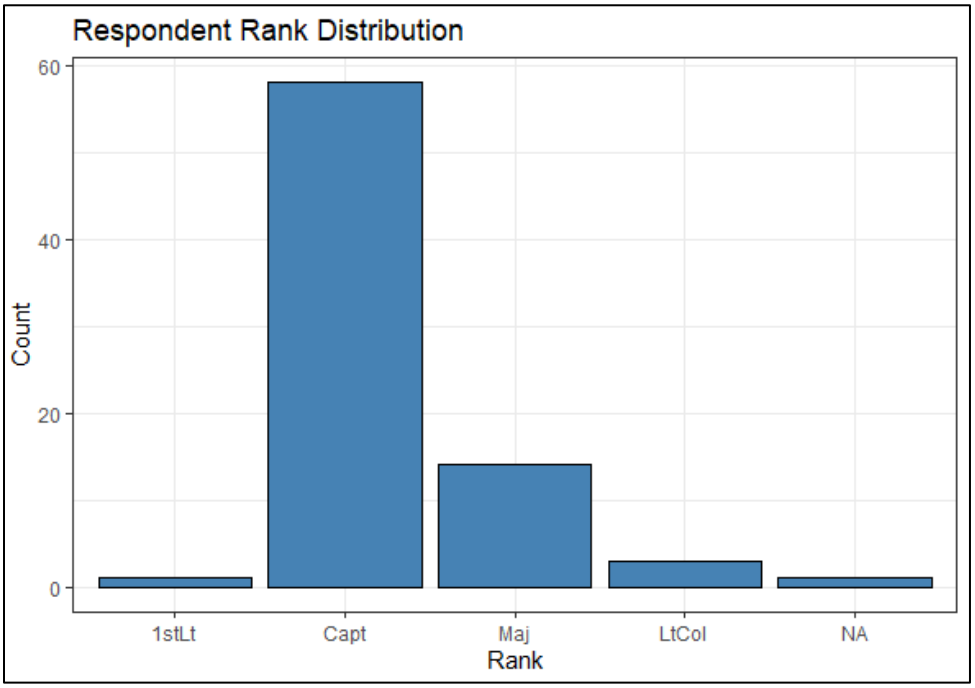


Figure 3. Bar chart depicting respondent rank demographic

The survey included two questions that allowed respondents to enter their four-digit MOS and select their MOS category. Refer to Figure 4 for a detailed breakdown of the

MOSs represented in the survey respondents, categorized by their respective four-digit MOS codes.

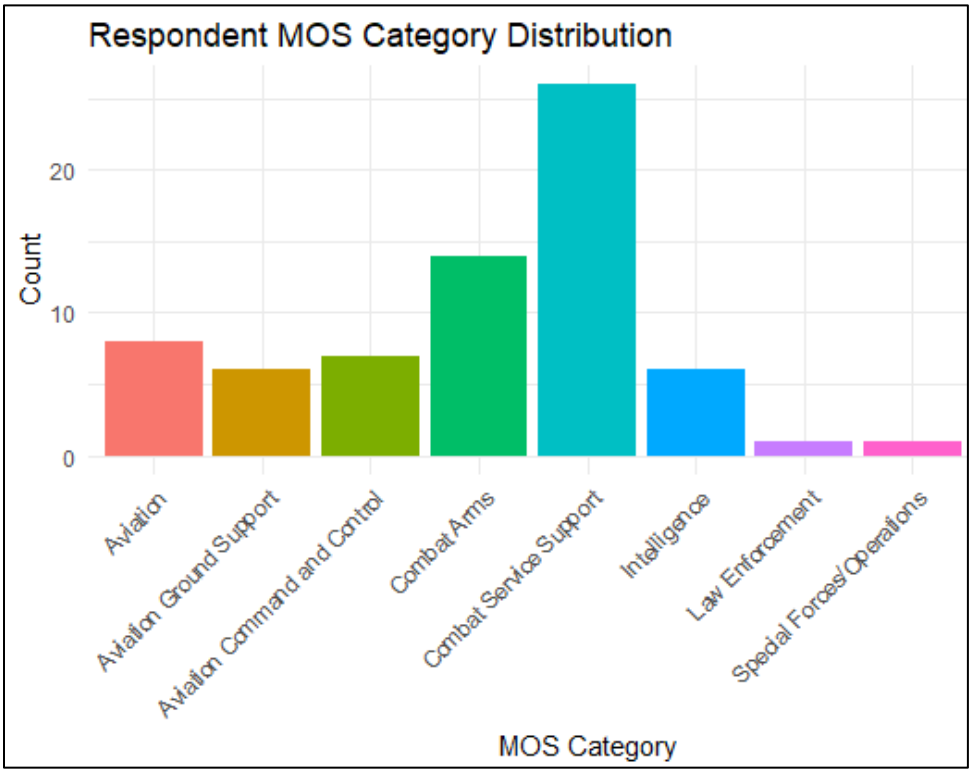


Figure 4. Bar chart depicting respondent demographics based on MOS category

Respondents were asked about their most recent training related to the PES, whether formal or informal This inquiry sought to address Clemens et al.'s (2012) assertion that a lack of report training contributes to disparities in performance evaluations and RS ratings. The analysis of training experiences and calculated inter-rater reliability aimed to determine whether there is support or opposition to the proposed solution suggested by Clemens et al. (2012) and Dunst (2018)—that additional training is necessary to minimize unexplained rater variations and potential bias.

A high level of reliability theoretically may indicate a unanimous understanding of PES concepts, attribute definitions, and attributed behavior, suggesting that training may

not be the primary issue, as implied by Puffpaff et al.'s (2015) findings, which reported no change in rater reliability pre and post training.

Among the respondents, 29 individuals reported having undergone only entry-level training on fitness report evaluations and writing at The Basic School (TBS). In contrast, a substantial portion, comprising 32 respondents, indicated that they had received additional command-level training beyond the basic curriculum at TBS. Additionally, 14 participants acknowledged having undergone some form of informal or formal training distinct from what was received at TBS. One person did not provide a response to this question. Refer to Figure 5 for a visual representation of the distribution of respondent training levels.

Figure 5. Bar chart depicting respondent demographics based on the highest form of training (formal or informal) received on fitness report evaluations and writing

Respondents were asked about the number of individual Marines they had assessed, aiming to gauge their experience in conducting performance evaluations and determine if they had an RS profile, which requires completing three individual assessments per grade. It is important to note that this inquiry pertained to the number of Marines evaluated, not

the number of reports written. The results revealed that most respondents had conducted more evaluations on enlisted Marines ranked E5 to E7 than any other rank. A sizeable proportion (38 out 76 respondents) had assessed senior enlisted Marines ranked E8 to E9, and a smaller subset had evaluated company-grade officers, including the rank of the MRO in the survey performance evaluation. See Figure 6 for visual distribution of Marines written on.

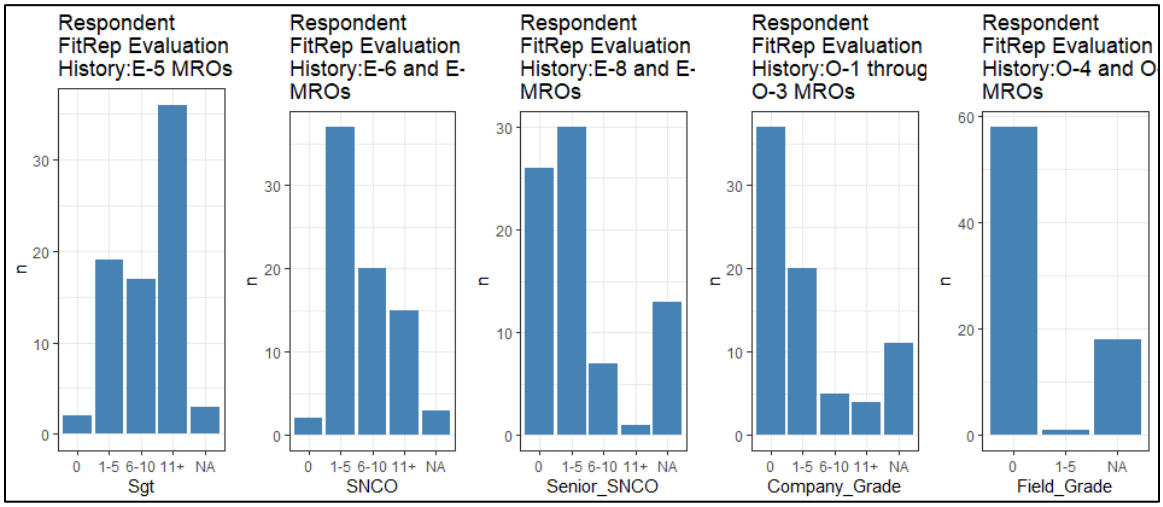


Figure 6. Bar chart depicting distribution of individual Marines assessed by respondents, throughout their career. The x-axis represents the number of respondents while the y-axis indicates the rank groups of individual Marines assessed.

B. INTER-RATER RELIABILITY

Our initial analysis of inter-rater agreement and consistency was from a descriptive statistical perspective, reviewing the report mean of each respondent and differences between respondent demographic categories to identify potential correlations that would address the secondary research question regarding relationships between PARS ratings and respondent demographics.

1. Quantitative Analysis of PARS Ratings and Averages

Of the 76 respondents, 51 completed the FitRep evaluation of the 8 attributes using the PARS. Respondent demographics remained relatively diverse in terms of rank.

Notably, a majority of respondents were Captains (37), with additional representation from various ranks, including Majors and Lieutenant Colonels. All MOS categories were represented. Additionally, respondents represented all levels of FitRep training backgrounds, with 15 having received entry-level training at TBS, 22 undergoing command-level training, and 13 participating in other forms of training.

It is essential to consider the PARS rating scale, where A is typically reserved for adverse material, B meets expectations typically associate with average performance, C-D are for consistent quality or exceeding expectations usually found with above average reports, E-G are for exemplary performance, and H is not observed or cannot be rated without observation. The distribution of scores across attributes suggests that respondents believed the MRO's overall performance ranged from commendable to above average, with instances of both average and exemplary competence in specific attributes. The most prevalent rating for the MRO was a C, as illustrated in Figure 7, which provides a visual distribution of attribute ratings, and Table 6, which presents the numeric count of ratings.

The MRO did not receive any adverse ratings. However, specific attributes, including Performance (54%), Initiative (50%), Setting the Example (45%), and Decision-making Ability (43%), received a high proportion of D ratings. Additionally, Judgment (54%), Courage (45%), Setting the Example (45%), Communication (45%), and PME (43%) received a high proportion of C ratings. Performance was the only attribute without any H ratings.

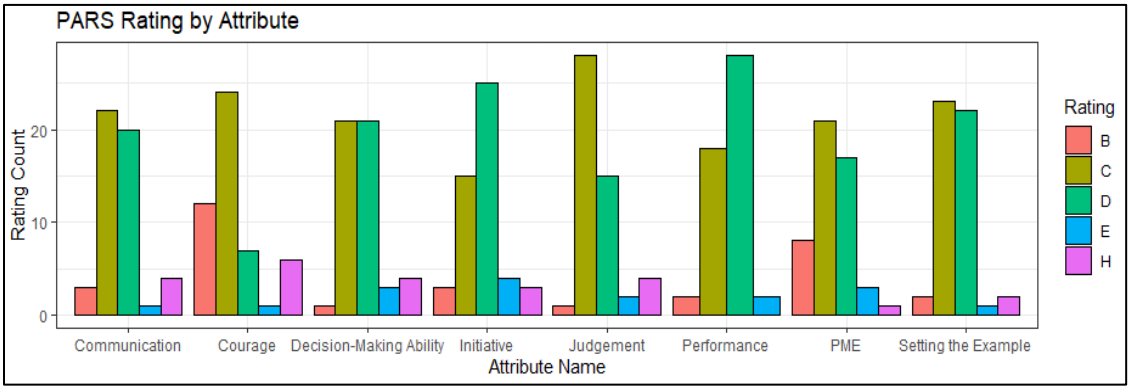


Figure 7. Visual illustration of PARS rating for each of the eight evaluated attributes

Table 6. Numerical depiction of PARS rating for each of the eight evaluated attributes

ATTRIBUTE	PERFORMANCE ATTRIBUTE RATING SCALE							
	A	B	C	D	E	F	G	H
Performance	0	2	18	28	3	0	0	0
Courage	0	12	25	7	1	0	0	6
Initiative	0	3	15	25	4	0	0	3
Setting the Example	0	2	23	23	1	0	0	2
Communication	0	3	23	20	1	0	0	4
PME	0	8	22	17	3	0	0	1
Decision-Making Ability	0	1	21	22	3	0	0	4
Judgement	0	1	28	16	2	0	0	4

Report averages were computed by assigning numeric values to each rating score, where "A" corresponds to 1 and "G" to 7. To maintain consistency and accurately reflect the fact that the "H" ratings could not be assessed, these were replaced with "NA," thus excluding them from the average calculation. This approach aligns with the understanding that an "H" rating signifies the RS deems the attribute unable to be assessed, and, consequently, the MRO does not receive a grade or points for that specific attribute. The alternative, not accounting for "H" values in the average, could potentially lead to inflated evaluations where the RS, despite rating multiple attributes as "H," provides a high overall grade since only "A" through "G" grades are considered. For instance, if a respondent rates seven attributes with an "H" and one attribute with a "C," the report average would be 3.00, as illustrated in Figure 8, depicting the distribution of report averages for all respondents.

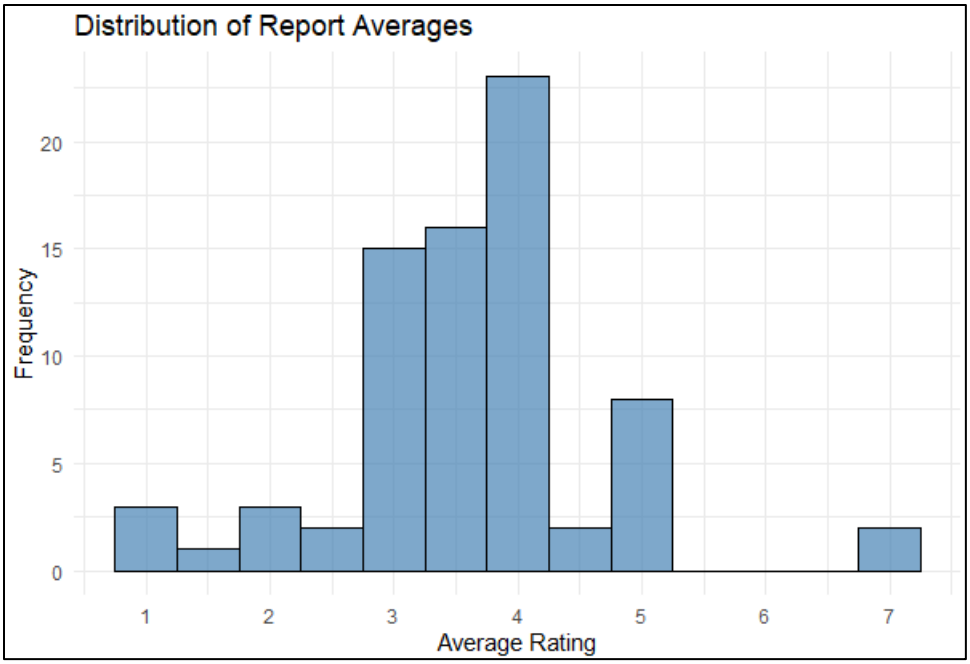


Figure 8. Bar chart depicting distribution of calculated report averages based on respondent PARS ratings

To discern any potential patterns or significance, we analyzed the averages based on training experiences, rank, and MOS category. Among the respondents who completed the evaluation, 72% had undergone additional training post TBS. We assessed the inter-rater reliability of both the TBS entry-level training group and the combined group for command-level and other training forms and found agreement to be low among and between both groups. Specifically, the Kappa score for TBS was 0.0387, and for other training, it was 0.046. Both scores yielded a p-value of 0.143, indicating a lack of statistical significance. See Figure 9 for cluster chart of report averages by training experience category.

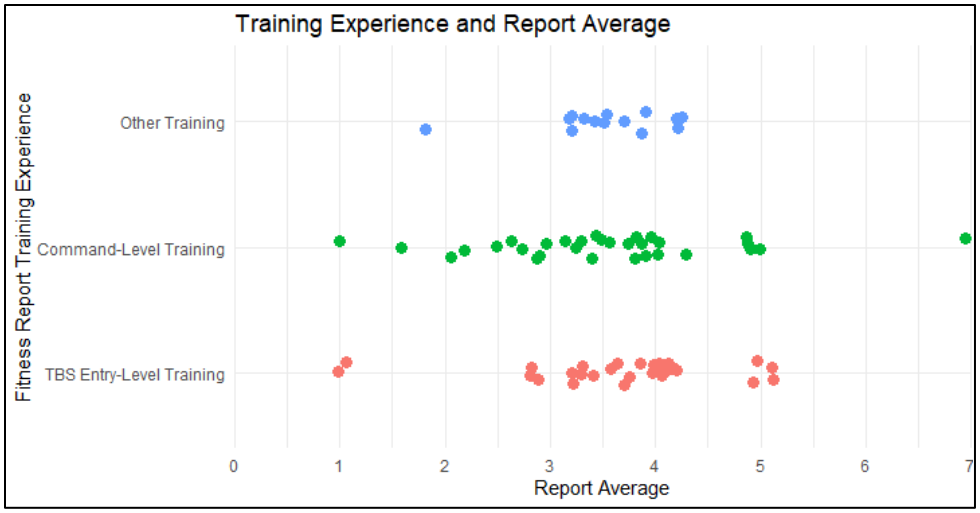


Figure 9. Cluster chart depicting distribution of report averages categorized by respondent FitRep training level

In further examining training experiences, we observed slight differences in mean scores of each group. Respondents with other training experiences had mean scores estimated to be 0.22 points higher than those with command-level training, while respondents who underwent TBS entry-level training had mean scores estimated to be 0.21 points higher than those with command-level training. Additionally, there was a negligible difference of 0.01 points favoring other training over TBS entry-level training. It is possible but not likely that these observed differences would have a significant impact on the MRO such moving them from the number 2 to number 1 position on the RS profile. Despite these observed variations and the lack of statistical significance, within the scope of this study, differences in training backgrounds do not exert a substantial influence on the performance evaluations of Marines.

Figure 10 visually depicts the distribution of report averages by rank. The box plot effectively illustrates the overall range of report averages as indicated by the blue box, their central tendency as indicated by the black line, and facilitates comparison of score distributions across the various demographic groups analyzed. Lieutenant Colonels collectively exhibited the lowest report averages. However, it's imperative to interpret this observation cautiously due to the limited representation of this rank category, which comprised the least number of respondents of all ranks.

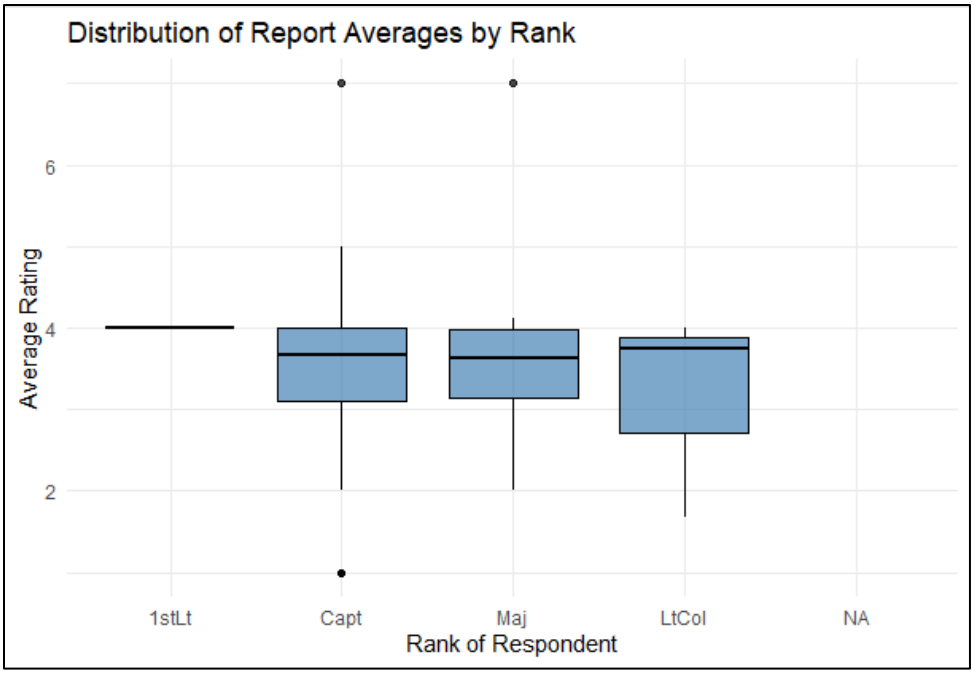


Figure 10. Box plot demonstrating report averages aggregated by rank of respondent. Black dots represent outliers.

The analysis of report averages by MOS category yielded a p-value of 0.0834, indicating that there may not be a statistically significant difference in mean ratings across different MOS categories at the conventional significance level of 0.05. However, it's worth noting that the result is close to the threshold, suggesting the need for further investigation. See Figure 11 report laying out the average organized by MOS category.

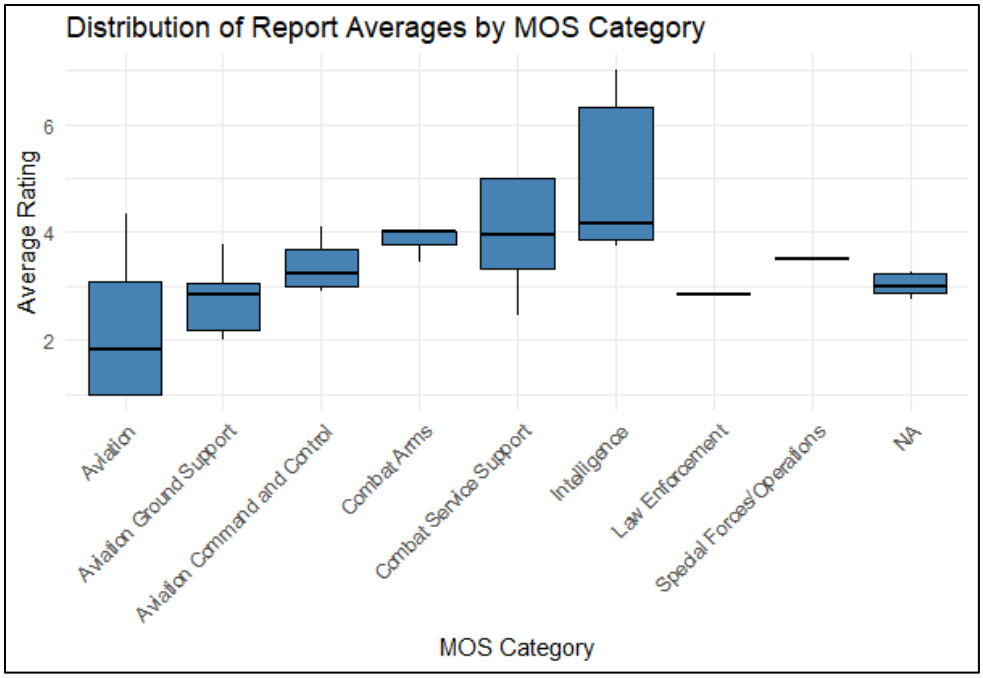


Figure 11. Boxplot demonstrating report average aggregated by MOS category

It was anticipated that MOS categories such as Aviation and Combat Arms would exhibit high variance due to the inherent differences between these communities. Aviators and Combat Arms personnel, based on the nature of their billets and locations, may have limited interactions with Ground Supply Officers, potentially resulting in a lack of familiarity with their roles and responsibilities compared to those within the Combat Service Support community, to which Ground Supply Officers belong. We conducted a pairwise comparison to identify statistically significance differences between categories. The comparison between Combat Service Support and Aviation Command and Control indicated that Combat Service Support had a lower average report rating, with a difference of -0.188 and a p-value of 0.9998. However, the relatively high p-values, exceeding the conventional significance level of 0.05, suggested that the observed differences in average report ratings between Combat Service Support and the other MOS categories lack statistical significance.

To test our theory on the variance in Combat Service Support community, given its closeness to the Ground Supply MOS, we isolated the category and performed a one sample

t-test to analyze the variance within this category (without reference to an external standard or benchmark). See Figure 12 for range of report averages (low 2.125 and high 4.25). The one-sample t-test conducted on the report average score ranges within Combat Service Support revealed a remarkably small p-value of 1.14×10^{-15}, indicating a highly significant result. Although unexpected, the significant variation could be the result of additional scrutiny applied by other Ground Supply Officers versus Financial Managers, and Logistics and Manpower officers.

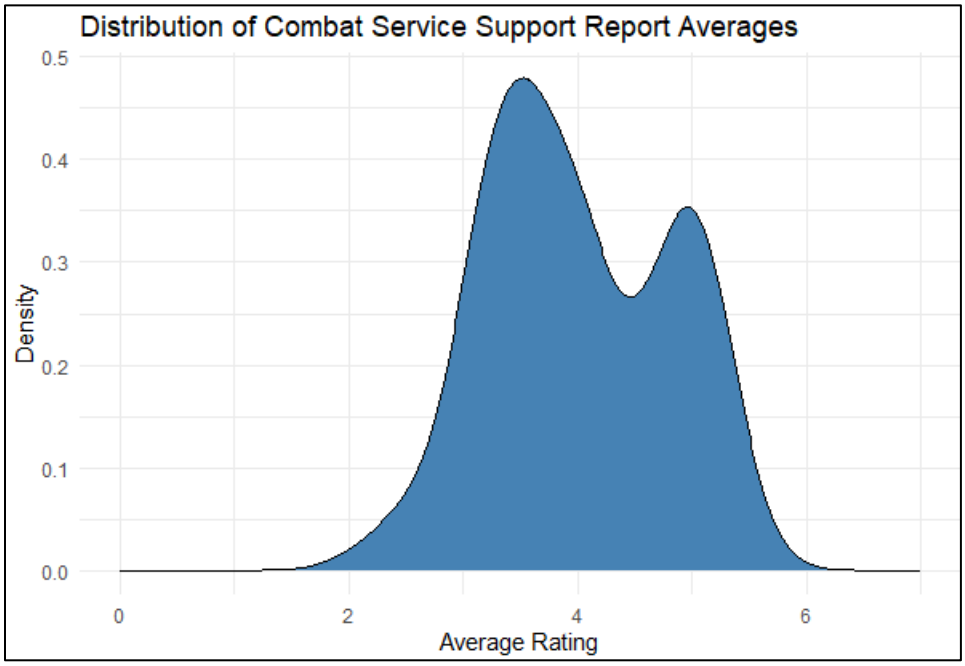


Figure 12. Density graph depicting report average for respondents in the Combat Service Support MOS category

Respondents were asked to provide an overall assessment rating for the MRO's performance based on their RS profile or expectations for a Marine of the of the MRO's rank, selecting from options that included "below average," "average" and "above average." The mean report average was 3.92 for Above Average, 3.39 for Average, and 2.82 for Below Average. As expected, the visual representation in Figure 13 reflects a trend where below-average assessments have the lowest scores, average assessments fall in the middle, and above-average assessments at the top. Notably, there was a surprising range of

scores within each assessment category. The range of report averages for the "average" assessment group alone spanned from 2.00 to 4.75.

To measure this variability, standard deviation and score range were calculated. The scores varied by 0.5 for above average (sd 0.188), 2.75 for average (sd 0.484), and 1.12 for below average (sd 0.438), providing insight into the degree of variability within each group.

Additionally, we conducted an ANOVA test. The ANOVA test assessed whether there were statistically significant differences in the mean values across the three categories for the overall assessment rating. The resulting F value was 7.766, and the associated p-value (Pr(>F)) was 0.0012. These results indicate that the differences between the means are statistically significant.

Further analysis revealed specific differences between the assessment categories. The report mean for the "Average" group is estimated to be 0.5302 points lower than the report mean for the "Above Average" group, a statistically significant difference ($p = 0.0292$). Similarly, the report mean for the "Below Average" group is estimated to be 1.0917 points lower than the mean for the "Above Average" group, with a highly significant difference ($p = 0.0007701$). Lastly, the report mean for the "Below Average" group is estimated to be 0.5615 points lower than the mean for the "Average" group, also statistically significant ($p = 0.0338$).

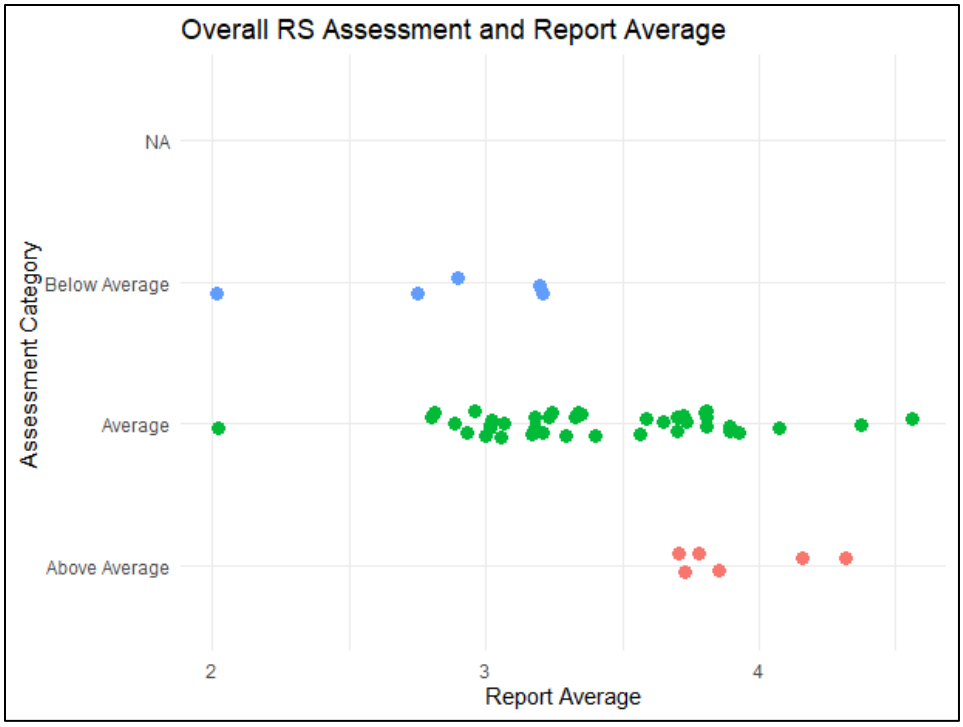


Figure 13. Cluster chart depicting distribution of report averages categorized by the overall assessment category for the MRO's performance

The takeaway from the analysis is despite respondents having the same level of information, observation, and grading scale, there is a notable variance in scoring which suggests that respondents interpret and apply the grading criteria differently, leading to diverse evaluations of the MRO's performance.

This variance could be influenced by individual biases, differing perspectives on what constitutes "Above Average," "Average," or "Below Average" performance, and variations in personal standards. It emphasizes the subjectivity inherent in performance evaluations, even when using a standardized grading scale. This insight is crucial for understanding the limitations and potential subjectivity associated with performance assessments, prompting organizations to consider additional measures, such as calibration sessions or clearer communication of grading criteria, to enhance consistency in evaluations.

2. Fleiss' Kappa

The Fleiss' kappa statistic and Light's kappa are valuable tools, especially in scenarios where multiple raters evaluate subjects or attributes using categorical scales with three or more categories (Mitani et al., 2017). Recognizing the categorical nature of the PARs scale, we treated it as a qualitative scale with predefined categories, such as 'adverse,' 'meets expectations,' 'above average,' and 'excellent.' Here, respondents rated the MRO based on these qualitative descriptors, rather than providing numerical scores for quality of attribute standards.

We removed all observations containing a "H" rating to better identify agreement amongst completed assessments. Upon computing Fleiss' Kappa, we obtained a value of 0.0261 with a corresponding p-value of 0.00234. This outcome suggests slight agreement among the 44 raters across various attributes. Importantly, the low p-value indicates that the observed agreement level is statistically significant, strongly suggesting that it is unlikely to have arisen by chance alone. Light's Kappa, albeit slightly higher at 0.054, also highlights the challenge of achieving consensus among raters.

These findings shed light on significant discrepancies in attribute assessments using the PARS rating system. Several factors may contribute to this lack of agreement, including differences in how raters interpret attribute descriptions, subjective biases, or inconsistencies in applying the rating criteria. Addressing these issues could enhance the reliability and validity of the assessment process, ensuring more consistent and accurate evaluations of performance.

3. ICC

We calculated the ICC to evaluate both the agreement and consistency of average scores from an ordinal perspective. In this context, we viewed the PARS utilized for attribute assessment as akin to a Likert scale, where 'B' corresponds to slight satisfaction with the MRO's performance, and 'G' represents extreme satisfaction. This interpretation assumes an ordinal nature, with performance being graded along a sliding scale, in contrast to the qualitative descriptors used in Fleiss' Kappa analysis.

The ICC for average score agreement yielded a value of 0.824 (95% CI: 0.624 - 0.955), indicating substantial agreement among raters regarding the assessment of attributes. This suggests that raters generally agree on the performance levels of Marine MRO based on the given scale, despite subjectivity in interpretation. Similarly, the ICC for consistency produced a value of 0.897 (95% CI: 0.761 - 0.975), signifying a high level of consistency among raters in their evaluations.

While the evaluation of performance using the PARS ratings is based on qualitative factors, these ratings are transformed into a numerical hierarchy when calculating averages. This dual nature of the rating system generated the conflicting results.

The discrepancy between the results of Fleiss' and Light's Kappa versus the ICC may indicate that while raters may not fully agree on individual attribute assessments, they tend to agree on the overall performance levels of the MRO when only considering average scores. Addressing factors such as differences in interpretation, biases, and inconsistencies in applying rating criteria could improve the reliability and validity of the assessment process, ensuring more consistent and accurate qualitative evaluations of performance.

C. TEXT ANALYSIS

Of the 51 respondents that completed the evaluation, 44 completed the justification narrative. Analysis revealed insights into the factors influencing the PARS ratings for the MRO, reflecting diverse perspectives on performance, attributes, and challenges in an evaluation based solely on the MROW.

Initial analysis was conducted using the R "tm" package to clean and filter the data and create a word cloud. This method was abandoned when the final product produced appeared convoluted and did not portray tangible or valuable findings. See Figure 14 for the Performance justification word cloud.

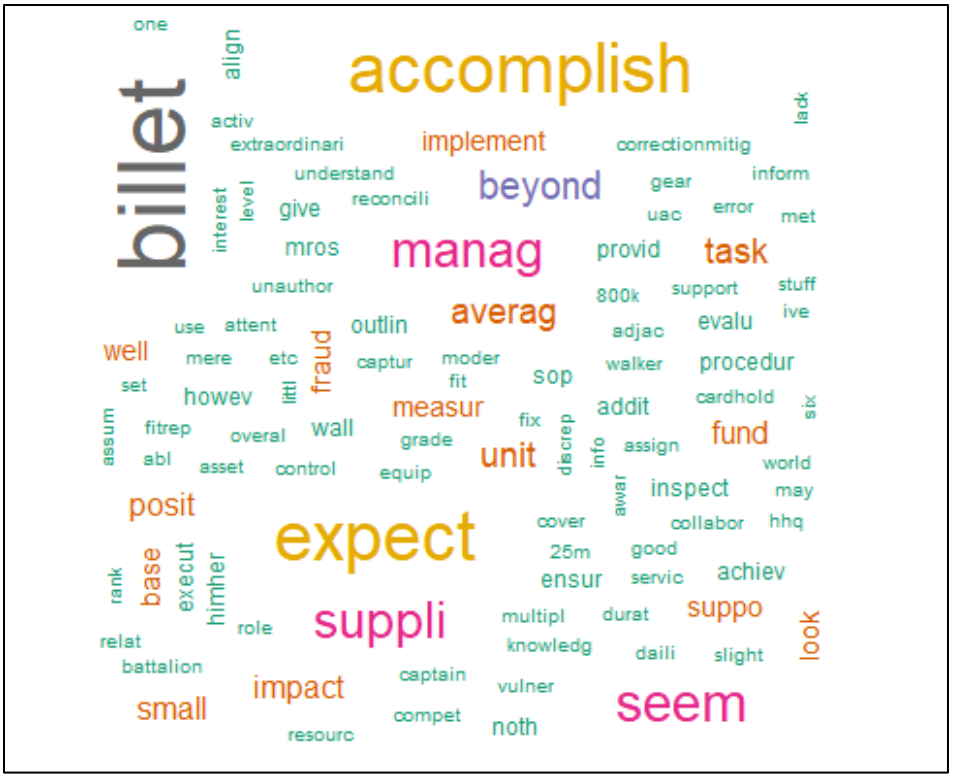


Figure 14. Word cloud image for performance justifications produced using R "tm" package

In examining the first attribute, Performance, it becomes apparent that there was a general consensus among respondents regarding the MRO's execution of duties as defined in MCO 1610.7B. Most respondents agreed that the MRO met billet expectations, emphasizing accountability and process improvements. However, when delving into the distribution of PARS ratings, some interesting patterns emerge:

1. B Ratings: The majority of respondents rated the MRO as meeting billet requirements, with a specific mention of their handling of small-arms inventories. This indicates a solid performance in meeting the basic expectations of the role.

2. C Ratings: Some respondents perceived the MRO's performance as average, expressing a desire for more impactful accomplishments. This suggests a divergence in expectations between respondents and highlights areas where the MRO could potentially improve to meet higher standards.

3. D Ratings: A subset of respondents acknowledged the MRO's successful management of funds and equipment but noted limited impact beyond expectations. This indicates a recognition of competency in core duties but a lack of significant contributions that go above and beyond.

4. E Ratings: Few respondents attributed proactive process improvements, significant budget management, and personnel management to the MRO. This suggests exceptional performance, exceeding billet expectations and demonstrating initiative and leadership beyond the norm.

While there was a general consensus on the MRO meeting billet expectations, there were variations in the assessment to which the MRO excelled or fell short. This indicates the subjective nature of performance evaluation and highlights the importance of clear criteria and communication in assessing performance attributes. Additionally, the need for more information about the Marine's daily tasks and performance outside of provided accomplishments highlights the challenge of evaluating performance comprehensively based solely on written reports.

In reviewing the assessment of the second attribute, Courage, it's evident that respondents faced challenges in evaluating this trait solely based on written reports. While the criteria for courage encompassed physical and moral strength, responsibility, and decision-making, respondents struggled to discern extraordinary acts of courage from routine tasks or accomplishments. Here's a breakdown of the distribution of ratings and some reflections:

1. B Ratings: Respondents cited limited information as a barrier to assessing courage, highlighting the need for face-to-face interaction to truly gauge this attribute. This suggests a consensus among respondents regarding the difficulty of evaluating courage solely through written reports.

2. C Ratings: Some respondents acknowledged moral courage in financial management, indicating a recognition of courage in handling fiscal

responsibilities. This implies a divergence in perspectives on what constitutes courageous behavior within the context of the MRO role.

3. D Ratings: Recognition of assertive interactions and courage in a staff billet suggests that some respondents identified instances of courage beyond routine tasks. However, the distribution of these ratings raises questions about the consistency of criteria applied across respondents.

4. H Ratings: The difficulty in assessing courage through the MROW alone was a common theme, indicating a consensus among respondents on the limitations of written reports in capturing courageous acts.

Overall, the distribution of ratings underscores the challenges in evaluating courage based solely on written reports, with respondents recognizing the limitations of the MROW in providing a comprehensive understanding of this attribute. However, there appears to be variability in the interpretation of courageous behavior, raising questions about the consistency of criteria application and the need for clearer guidelines in assessing this trait.

The assessment of the third attribute, Initiative, reveals a spectrum of feedback ranging from average to above-average observations of initiatives demonstrated by the MROs. While respondents generally recognized proactive steps taken by the MROs to improve processes, conduct internal inspections, and establish SOPs, there were differing opinions on whether these actions constituted routine tasks or extraordinary initiatives. Here's a summary of the rating distributions and some reflections:

1. B Ratings: Some respondents noted a lack of observative initiative beyond routine tasks, indicating a desire for more impactful demonstrations of initiative.

2. C Ratings: Acknowledgment of routine accomplishments was observed, with respondents expressing a need for more tangible evidence to assess the level of initiative accurately.

3. D Ratings: There was a mixed but generally positive perception of demonstrated initiative, particularly in tasks such as inventory accountability and SOP implementation. However, respondents sought more clarity on the distinction between routine and extraordinary initiatives.

4. E Ratings: Respondents highlighted proactive engagement in self-directed internal inspections and the development of programs as clear demonstrations of initiative. These responses emphasize the need for impactful actions that align with the commander's intent.

Overall, the responses underscore a desire for MROs to showcase independent action that goes beyond routine tasks, with a focus on proactive and impactful initiatives that align with organizational goals.

The fourth attribute, Setting the Example, encompasses how effectively a Marine embodies ethical behavior, fitness, appearance, bearing, and self-discipline, serving as a role model for others. Physical fitness scores emerged as the primary metric cited by respondents to evaluate this attribute. Many acknowledged positive aspects of the MRO's performance, such as consistently achieving first-class PFT/CFT scores, maintaining height and weight standards, and attaining expert marksman status. Additionally, respondents recognized the MRO's commitment to duty through engagement in training, mentoring, and participation in events like Regional, Culture and Language Familiarization (RCLF) or supply symposiums. However, there was a recurring emphasis on the significance of personal knowledge and direct observation in accurately assessing whether the Marine sets a commendable example for peers and subordinates. Here's a summary of the rating distributions:

1. B Ratings: Respondents noted insufficient information to justify a higher rating, expressing a need for more details on mentoring and training activities.

2. C Ratings: The MRO met fitness standards and displayed average performance, but there was limited evidence of going above and beyond expectations.

3. D Ratings: Ratings in this category relied on metrics like dedication to duty, leadership, and engagement in mentorship activities.

4. E Ratings: The MRO demonstrated solid performance in physical training, and actively trained and mentored a squad-sized element, earning high praise in this regard.

These ratings underscore the importance of not only meeting basic requirements but also demonstrating proactive engagement and leadership qualities to set a positive example for others.

The fifth attribute was Communication. While the criteria for communication encompassed strong listening, speaking, writing, and critical reading skills, respondents faced challenges in discerning the effectiveness of communication solely from the provided information. Here's a breakdown of the distribution of ratings and some reflections:

1. B Ratings: Respondents called for more concrete examples to substantiate higher ratings.

2. C Ratings: Some respondents expressed challenges in assessing communication due to limited information, indicating assumptions of average communication skills based on the provided data (i.e., diverse account management. This suggests a divergence in perspectives on what constitutes effective communication within the context of the MRO role.

3. D Ratings: Recognition of coordinating funds and effective account management suggests that some respondents identified instances of effective communication demonstrated by the MRO. However, the distribution of these ratings raises questions about the consistency of criteria application across respondents.

4. E Ratings: Emphasis on clear articulation, attention to formatting, and grammar within the MROW indicates a recognition of the importance of these aspects in effective communication.

The distribution of ratings suggests varying interpretations of what constitutes exemplary communication skills. Respondent identified additional information that would have proved helpful in assessing this attribute which included direct observation of verbal communication, and examples of written communication, such as emails, letters of instruction, point papers, and standard operating procedures.

In examining the assessment of the sixth attribute, Professional Military Education (PME), it's evident that respondents held varied perspectives on this trait, reflecting differing opinions on the significance of additional PME activities. While some respondents viewed completion of activities such as RCLF and attendance at supply symposiums as indicative of above-average performance, others emphasized the importance of enrollment or completion of required PME, particularly Expeditionary Warfare School (EWS), in their assessments. Here's a breakdown of the summary ratings:

1. B Ratings: Respondents noted the absence of enrollment in EWS but recognized completion of other PME activities such as RCLF. This suggests a consensus among respondents regarding the importance of pursuing additional PME beyond standard requirements.

2. C Ratings: Some respondents acknowledged the completion of required PME but expressed uncertainty regarding EWS enrollment. There were expectations for efforts beyond grade-specific requirements, indicating a divergence in perspectives on the adequacy of PME efforts.

3. D Ratings: Recognition of additional PME efforts was noted, although there remained uncertainty regarding EWS status. This suggests variability in the interpretation of PME requirements and the importance placed on specific educational activities.

4. E Ratings: Active participation in conferences and completion of RCLF were highlighted as evidence of a proactive approach to professional development within the MOS. This indicates a recognition of the importance of engaging in diverse PME opportunities to enhance professional growth.

Overall, the results underscore the importance of ongoing professional development through PME activities beyond standard requirements. While there is recognition of various PME efforts, there is also a need for clarity and consistency in assessing the fulfillment of PME requirements, particularly regarding enrollment in essential programs such as EWS. Additionally, the emphasis on pursuing a range of military and civilian educational opportunities reflects a broader understanding of the value of continuous learning and intellectual development in the Marine Corps.

In reviewing the assessment of the seventh attribute, Decision-Making Ability, it's apparent that respondents generally expressed positivity towards this trait, highlighting effective problem-solving and balanced judgment. Indications of effective and efficient decision-making were said to be demonstrated through successful management of tasks, achievement of billet accomplishments, and proactive measures to identify and address vulnerabilities. The Marine was perceived to handle responsibilities well, make sound decisions, and manage resources effectively within the given billet. Here's a breakdown of the summary ratings:

1. B through D Ratings: Respondents noted the difficulty in conclusively assessing decision-making ability without witnessing firsthand. This suggests a consensus among respondents regarding the importance of direct observation and the challenges in accurately evaluating decision-making ability without additional context.

2. E Ratings: The Marine was perceived to go beyond the scope of the billet description, demonstrating proactive measures to identify and address vulnerabilities. This indicates a recognition of exceptional decision-making ability and proactive leadership in managing tasks and challenges.

Overall, the results reflect a generally positive sentiment towards the Marine's decision-making ability, with acknowledgments of the challenges in assessing this trait in a simulated scenario without direct observation.

In evaluating the eighth attribute, Judgment, respondents grappled with the challenge of assessing this trait without direct observation, leading to an overall sentiment marked as "Average." Here's a breakdown of the rating justifications:

1. B Ratings: Respondents noted that accomplishments appeared to be sound grounded in logic and fundamental teachings.

2. C Ratings: Some respondents highlighted the intertwined nature of judgment with decision-making ability indicating the difficulty involved in assessing judgment solely based on written reports.

3. D Ratings: Respondents noted that the absence of evidence showing superior or inferior judgment further challenged their ability to evaluate this trait through written reports alone.

4. E Ratings: The lack of observable instances of making logical and well-informed decisions led to lower ratings in this category.

Overall, the assessments reflect a nuanced understanding of judgment, with respondents grappling with the complexities of evaluating this trait solely based on written reports. Contextual information such as restraints, constraints, commander's priorities, and interaction or observation of the individual in various scenarios, especially challenging situations, to gauge their response and application of judgment were annotated as necessary to better evaluate this trait. The emphasis on context and observable instances of decision-making underscores the challenges and importance of assessing judgment accurately in a simulated scenario.

The analysis of justification narratives revealed a consensus among respondents regarding the difficulty of assessing attributes without sufficient contextual knowledge and direct observation. Despite all respondents having access to the same information and lacking direct observation, there was a notable variation in ratings. While some respondents

perceived the MRO's performance as 'average' or 'expected of an officer,' PARS ratings ranged from B to E. This variance in ratings suggests a nuanced interpretation of performance levels, potentially influenced by individual perspectives and critical thinking skills employed in the absence of observation.

Moreover, respondents demonstrated a heightened ability to synthesize information, draw meaningful conclusions, and make connections that were not immediately apparent. However, it's essential to consider whether this heightened critical thinking reflects the respondents' real FitRep evaluation practices with their subordinates. Further examination of the underlying factors influencing rating variations and critical thinking processes would provide deeper insights into the evaluation dynamics.

D. SUMMARY

The analysis provided a comprehensive examination of respondent statistics, including demographics, and training experiences, in addition to inter-rater reliability, providing valuable insights into performance evaluation agreement and consistency.

Respondent demographics revealed an overrepresentation of officers at the rank of Captain, grade (O3), and in MOS categories associated with Combat Service Support roles, such as Ground Supply, Financial Management, Manpower, and Logistics. Analysis of training experiences aimed to address discrepancies and scoring variations in performance evaluations. Our analysis revealed varying levels of training received among respondents, with the majority having undergone additional command-level training subsequent to their entry-level training at TBS. While some differences in mean scores across different training experiences were noted, the differences were not statistically significant.

Additionally, respondents reported on the number of individual Marines they had assessed, indicating a focus on evaluations of enlisted Marines ranked E5 to E7, with fewer assessments of senior enlisted Marines and company-grade officers. Most respondents had at least established an RS profile for multiple ranks.

Inter-rater reliability analysis explored the consistency of FitRep evaluations using the PARS. While there was general agreement on the distribution of attribute ratings, there

was variability in scoring, suggesting differing interpretations of grading criteria among respondents.

Insights from text analysis of justification narratives further underscore the challenges in assessing attributes without direct observation and the importance of contextual knowledge. Despite consistent information provided to respondents, ratings varied, reflecting subjective interpretations and critical thinking skills in evaluation processes.

Overall, the analysis highlights the intricate nature of performance evaluations, revealing varying interpretations of grading criteria among respondents. Moreover, it emphasizes the need for transparent communication of grading standards, and calibration sessions to aid in fostering consistency and minimizing biases in assessments. These findings draw attention to the importance of addressing the complexities of performance evaluations to ensure fairness and accuracy.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSION

A. DISCUSSION

Our findings suggest low inter-rater reliability amongst multiple raters from diverse backgrounds, in their limited evaluation of a single MRO, suggesting that the raters have difficulty applying the assessment criteria used in measuring performance consistently. This indicates potential issues with the validity and fairness of FitRep evaluations, as they may not reflect an absolute or balanced (subjective vs. objective) measure of performance and talent.

The small sample size and underrepresentation of the affected population limit generalization of findings. However, if the results were replicated on larger scale, it would call into question the FitRep ability to effectively identify and promote talent within the organization, by providing an accurate and fair assessment of Marines' performance. This shortcoming could undermine the integrity of the promotion and retention processes within the Marine Corps and may support the lack of trust in the system among personnel. The Marine Corps should apply additional resources to further investigate this issue and develop strategies for addressing the inconsistencies and unexplained variations in scoring.

The current study builds upon the existing body of research by specifically focusing on the reliability of FitRep evaluations providing empirical evidence to support concern and challenges within the Marine Corp PES.

Analysis of respondent FitRep training experience revealed no significant difference in the inter-rater reliability between groups, suggesting that additional training may not reduce unexplained variations in RS scoring as previously suggested by Dunst (2018) and Clemens et al. (2012). This finding, while not entirely aligned with research by Pufpaff et al. (2015) due to methodological differences (test-retest reliability), does call into questions the true value of providing additional training over modifying the absolute measures of the FitRep for enhanced standardization.

Studies conducted by the CNA and NPS students have raised concerns about unexplained score variations based on factors such as race, gender, MOS, and educational

background. While our study did not directly support or refute these findings, it aligns with previous research that has identified similar issues within the Marine Corps PES.

A comparison of text analysis and report averages revealed inconsistencies, particularly for Marines ranked as average overall, indicating discrepancies between the narrative suggesting the MRO is or is not meeting expectations and the assigned overall assessment score. These inconsistencies raise concerns about the ability of RSs to effectively apply the FitRep to differentiate between Marines of varying performance levels accurately. There are two key takeaways from this finding: one is the challenge of accurately assessing Marines ranked as average or below average using the PARS and two inflating the score of average Marines to make room for those not meeting expectations or displaying inconsistent performance. We found that PARS ratings in the average category ranged from B to D, indicating inconsistency in evaluating Marines of a similar performance level. This difficulty in distinguishing truly average Marines suggests that there may be a need to score them higher to allow room for those who are below average while not making the report adverse.

The text analysis revealed multiple attributes rated as "C" or "D" where respondents deemed the observable performance "average" or "expected" of an officer in that position. This finding suggests that Marines who are not meeting expectations but are not necessarily adverse are receiving "B" and "C" ratings which increase the report average and conflict with one of the intended purposes of the report which is to help differentiate personnel quality for the purpose of promotion, retention, and assignment.

The RV measure, intended to aid in distinguishing RS scoring trends by categorizing reports into thirds, would theoretically help create balance and alleviate the issue of inconsistent scoring. However, the effectiveness of RV in achieving this goal was called into question by Dr. Baker's 2024 report. Dr. Baker's findings demonstrate how this measure can be misleading, particularly in distinguishing between Marines ranked in the bottom and middle thirds of their RS profiles.

There is further alignment with Rigaut's 2017 study on section I and K comments. A disconnect between assessment narrative used to highlight performance and the report

average are more problematic for special assignments and field grade officer promotion boards which are more competitive and have a lower percentage selection rate. Again, here is where the RV is helpful to discern between top 10% and the truly "average" Marines on the margins. The opportunity to conduct the ideal experiment including repeated assessments by the same RSs would have enabled us to better observe the true impact of the disconnect by reviewing the word picture for different tier Marines and the overall assessment.

While performance evaluations inherently require a subjective component to assess behavior and character against organizational standards, maintaining a reasonable balance between objective and subjective aspects is crucial to ensure fairness. The PES Manual attribute guidance offers an absolute measure of performance, supporting the objective component, while the RS profile provides a relative measure, supporting the subjective aspect of the report. Despite this framework, challenges such as low inter-rater reliability, an imbalanced connection between narrative and PARS ratings, and the questionable effectiveness of the RV in deciphering relative assessments raise concerns about the reliability of the information presented to promotion boards. Furthermore, unexplained variances in MRO report averages by specific demographics add to the uncertainty surrounding the decision-making process.

B. LIMITATIONS

While our study offers valuable insights into the FitRep evaluation process, it is essential to acknowledge several limitations that may impact the generalizability and comprehensiveness of our findings.

1. Self-Selection Bias

Respondents were not randomly selected but instead opted to take part in the research. The self-selection process led to an overrepresentation of some officer groups and a lack of others. Furthermore, the pool of respondents may have only included individuals who were more motivated, interested, or have specific perspectives on the FitRep process.

2. Sample Representatives

The respondents came from students enrolled at NPS, potentially limiting the study's representativeness to the broader population of interest, which includes individuals in operational fleet units. This sample introduces a potential bias, as graduate students may differ from those in operational fleet roles limiting the comprehensive examination of perspectives and insights.

More specifically, it's essential to acknowledge that respondents in this study may have been influenced by the academic environment and objectives associated with post-graduate education, introducing a potential bias that may not align with the experiences and perspectives of individuals serving in operational roles. For example, some rating justification narratives appeared to use more critical thinking skills than would be applied in an authentic RS–MRO relationship with direct observation. Respondents inferred assumed performance and behaviors based on the billet, the rank of the MRO and assumed relationships without having direct observation or conclusive evidence to support assumptions.

3. Sample Size

The DOD 2021 Demographics Profile of the Military Community confirms that the Marine Corps has 21,701 active-duty officers. The survey sample size constitutes less than 1% of the target population. Generalizing findings requires a representative sample, and the limited size of the survey group restricts the ability to draw conclusions about how officers, in general, may agree or disagree with the described attribute ratings and correlating behavioral examples.

4. Lack of Observation

The absence of direct observation limited the study's ability to fully capture the intricacies of the FitRep evaluation process. Multiple respondents annotated how the lack of direct observation prevented them from directly witnessing and analyzing the MRO's performance. Without direct observation, they could not ascertain character, true initiative,

non-verbal cues or contextual factors, that could have aided in assessing the MRO's performance against the FitRep attributes.

C. CONCLUSION

This study has provided valuable insights into the reliability of Marine Corps FitRep evaluations and its implications for promotion and assignment boards. Through a comprehensive analysis of inter-rater reliability among RSs, we aimed to address concerns surrounding scoring inconsistencies and the subjective nature of FitRep assessments. Thorough exploration of existing literature, empirical evidence, and qualitative insights have revealed several critical shortcomings within the FitRep system.

The existing FitRep framework exhibits inherent limitations that suggest modifications are needed that may help provide a more fair, unbiased, and comprehensive assessment of Marine performance. Concerns surrounding attribute ambiguity, validity issues, and subjective interpretation by RSs have consistently been highlighted in both academic research and after-action reports. Moreover, disparities in scores based on demographic factors such as race, gender, and MOS raise serious questions about the system's equity and objectivity.

The findings contribute to enhanced comprehension of the challenges and limitations inherent to the PES. Despite attempts to establish uniform evaluation criteria and foster objectivity, the findings indicate notable disparities in ratings and assessments among RSs. This inconsistency prompts inquiries into the reliability and validity of FitRep assessments and their ability to accurately discern talent and performance.

The implications of these findings extend beyond the evaluation to impact promotion and retention decisions within the Marine Corps. Inaccurate or biased assessments can undermine morale, hinder career progression, and erode trust in the evaluation process. Despite efforts to address these concerns, the lack of funding allocated for evaluating proposed changes indicates a systemic inertia in addressing the pressing need for reform within the PES. This inertia is compounded by a growing lack of faith within the ranks, as acknowledged by the 38th Commandant's Planning Guidance.

Previous quantitative analyses (Clemens et al., 2012; Dunst, 2018) have indicated a consistency issue with ratings, prompting this study to examine whether the evaluation form itself contributes to this problem. The results suggest a low reliability in the assessment process. Consequently, it is imperative for the Marine Corps to reassess the current evaluation criteria to ensure a more balanced approach between objective and subjective elements. This revision could entail refining the attributes being assessed to effectively capture both tangible accomplishments and intangible qualities. Moreover, it's essential to establish a fairer ratio between the subjective interpretations of these attributes and the objective absolute measure of attributes.

While the attributes serve as an absolute measure, the trends identified in the text analysis highlight a heavy reliance on RS interpretation of demonstrated behaviors, which may be influenced by bias and intent. This is further supported by Jobst and Palmer's (2005) survey which found that officers weigh each attribute differently based on their MOS. Therefore, it is imperative that alongside revising the evaluation criteria, comprehensive training programs are provided to both RSs and ROs. These programs should aim to educate them on the intended evaluation process (e.g., attribute clarity, RS profile development, effective writing techniques, comparative assessment distributions, RV calculations and interpretation, etc.), and strategies to mitigate biases, thus ensuring fair assessments across the board.

Moreover, it may be worthwhile to explore the potential efficacy of implementing competitive category evaluations for officers, as suggested by Jobst and Palmer (2005). These evaluations would allow RSs to highlight standout performers within specific MOS categories, rather than evaluating officers across all MOSs in the same grade. This approach could potentially provide a more merit-based approach capitalizing on individual strengths and foster a fairer evaluation process overall. In addition, it is better aligned to new talent management efforts at HQMC which aim to modify our personnel management system by highlighting and rewarding individual talent.

Furthermore, considering modifications to the PARS to allow for ratings that accurately reflect below-average, inconsistent, or subpar performance could enhance the evaluation process's accuracy and fairness. By implementing these changes, the Marine

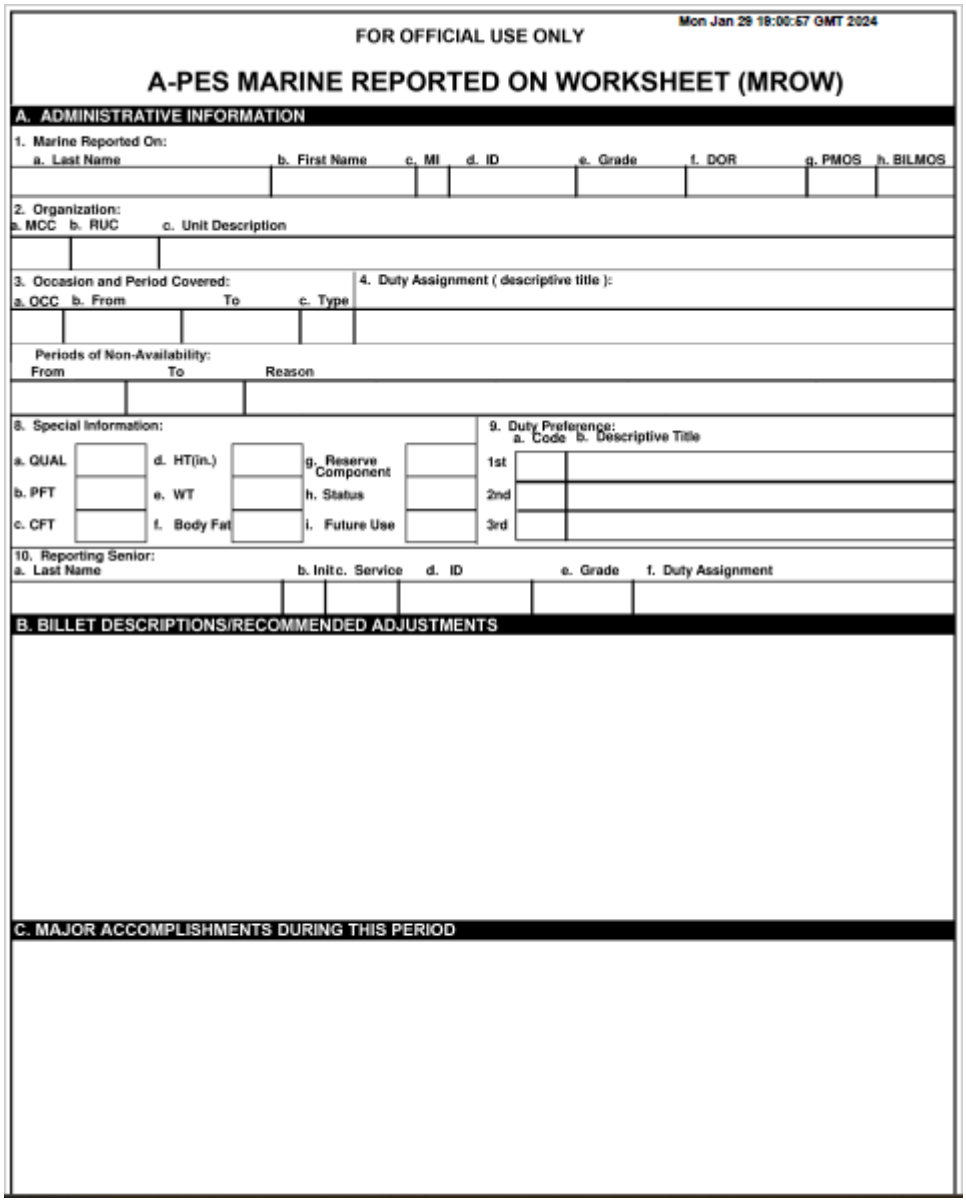
Corps can improve the integrity and effectiveness of its FitRep evaluations, ultimately better identifying and developing talent to fulfill its mission objectives.

Our current manpower system, developed during the industrial era, prioritizes quantity over quality, often overlooking individual talent and performance (Berger, 2019). To enhance the readiness and effectiveness of our force, it is essential to address the challenges outlined in this study through meaningful reforms. Moving forward, the Marine Corps must undertake systematic evaluation and reform of the PES. This includes revisiting performance assessment criteria, addressing subjective factors, and providing adequate training and support for Reporting Seniors (RSs). Additionally, expanding the scope of performance evaluation to encompass a broader range of skills and attributes aligned with organizational standards of quality is crucial.

By implementing these reforms, the Marine Corps can strengthen its ability to identify, develop, and retain top talent, thus ensuring readiness and effectiveness in fulfilling mission objectives.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. MARINE REPORTED-ON WORKSHEET



FOR OFFICIAL USE ONLY

Mon Jan 29 18:00:57 GMT 2024

A-PES MARINE REPORTED ON WORKSHEET (MROW)

A. ADMINISTRATIVE INFORMATION

1. Marine Reported On:
a. Last Name | b. First Name | c. MI | d. ID | e. Grade | f. DOR | g. PMOS | h. BILMOS

2. Organization:
a. MCC | b. RUC | c. Unit Description

3. Occasion and Period Covered:
a. OCC | b. From | To | c. Type

4. Duty Assignment (descriptive title):

Periods of Non-Availability:
From | To | Reason

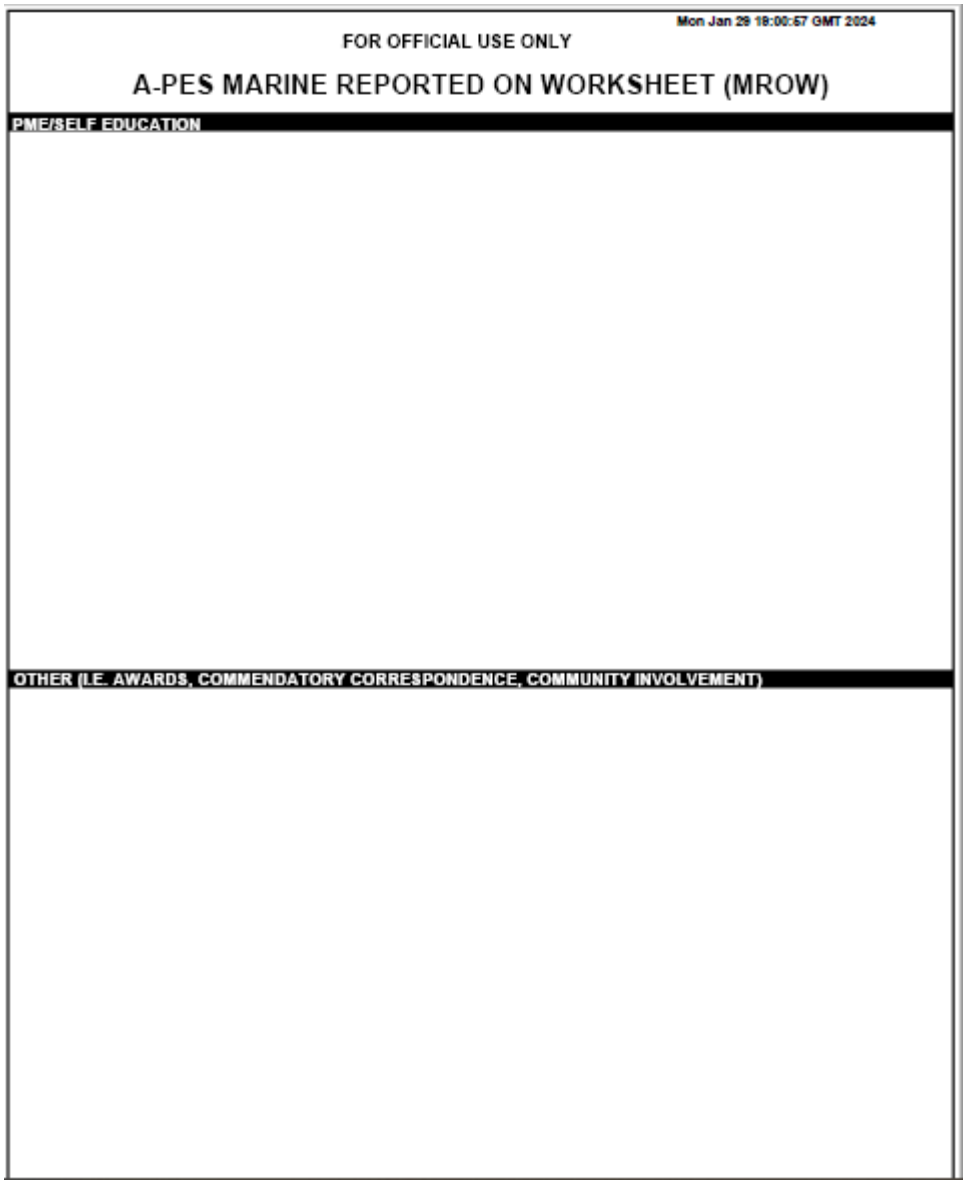
8. Special Information:
a. QUAL | d. HT(in.) | g. Reserve Component
b. PFT | e. WT | h. Status
c. CFT | f. Body Fat | i. Future Use

9. Duty Preference:
a. Code | b. Descriptive Title
1st
2nd
3rd

10. Reporting Senior:
a. Last Name | b. Init | c. Service | d. ID | e. Grade | f. Duty Assignment

B. BILLET DESCRIPTIONS/RECOMMENDED ADJUSTMENTS

C. MAJOR ACCOMPLISHMENTS DURING THIS PERIOD



Mon Jan 29 18:00:57 GMT 2024

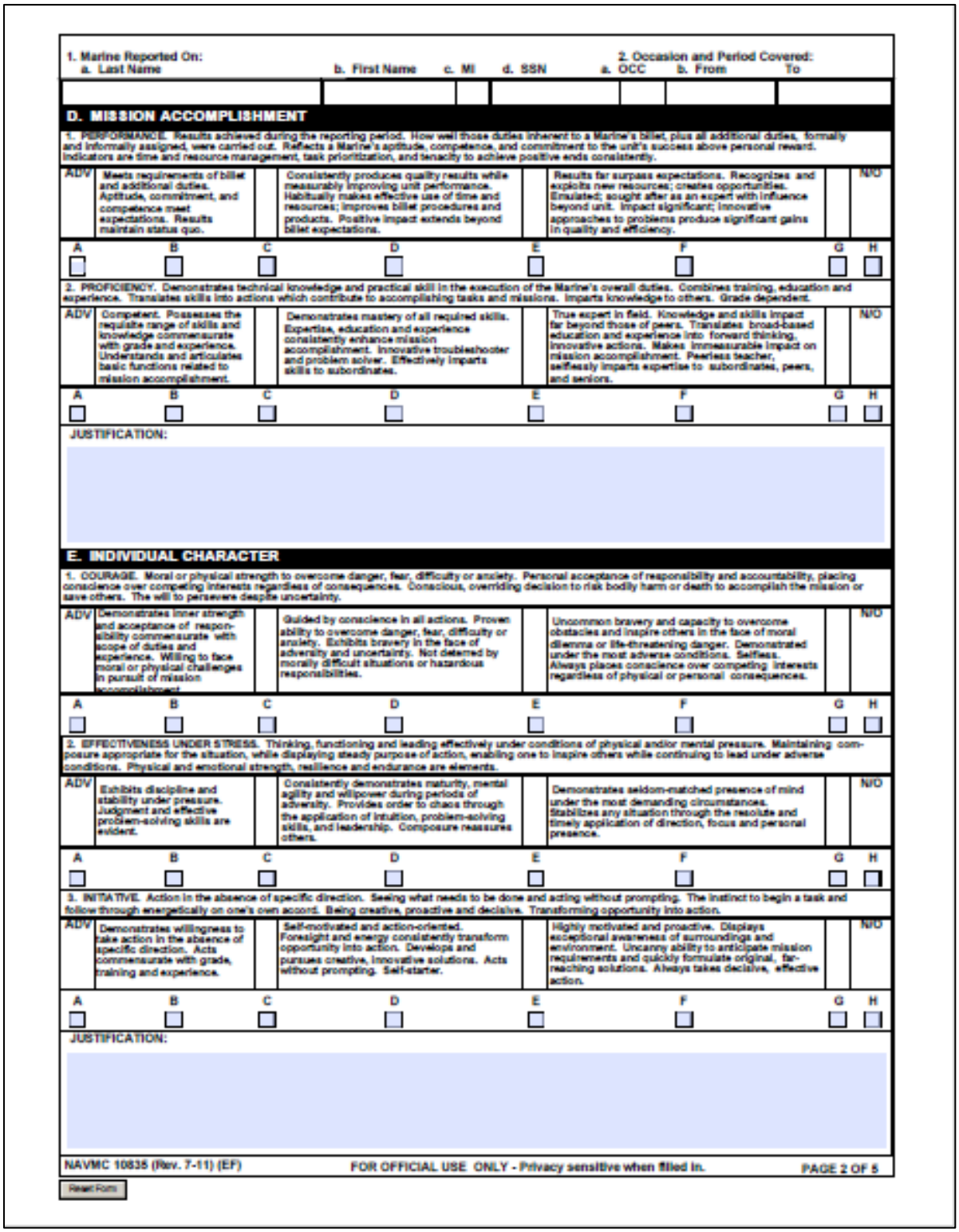
FOR OFFICIAL USE ONLY

A-PES MARINE REPORTED ON WORKSHEET (MROW)

PME/SELF EDUCATION

OTHER (I.E. AWARDS, COMMENDATORY CORRESPONDENCE, COMMUNITY INVOLVEMENT)

APPENDIX B. NAVMC 10835 (REV 7–11) FITNESS REPORT



1. Marine Reported On: a. Last Name | b. First Name | c. MI | d. SSN | 2. Occasion and Period Covered: a. OCC | b. From | To

D. MISSION ACCOMPLISHMENT

1. PERFORMANCE. Results achieved during the reporting period. How well those duties inherent to a Marine's billet, plus all additional duties, formally and informally assigned, were carried out. Reflects a Marine's aptitude, competence, and commitment to the unit's success above personal reward. Indicators are time and resource management, task prioritization, and tenacity to achieve positive ends consistently.

ADV			N/O
Meets requirements of billet and additional duties. Aptitude, commitment, and competence meet expectations. Results maintain status quo.	Consistently produces quality results while measurably improving unit performance. Habitually makes effective use of time and resources; improves billet procedures and products. Positive impact extends beyond billet expectations.	Results far surpass expectations. Recognizes and exploits new resources; creates opportunities. Emulated; sought after as an expert with influence beyond unit. Impact significant; innovative approaches to problems produce significant gains in quality and efficiency.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

2. PROFICIENCY. Demonstrates technical knowledge and practical skill in the execution of the Marine's overall duties. Combines training, education and experience. Translates skills into actions which contribute to accomplishing tasks and missions. Imparts knowledge to others. Grade dependent.

ADV			N/O
Competent. Possesses the requisite range of skills and knowledge commensurate with grade and experience. Understands and articulates basic functions related to mission accomplishment.	Demonstrates mastery of all required skills. Expertise, education and experience consistently enhance mission accomplishment. Innovative troubleshooter and problem solver. Effectively imparts skills to subordinates.	True expert in field. Knowledge and skills impact far beyond those of peers. Translates broad-based education and experience into forward thinking, innovative actions. Makes immeasurable impact on mission accomplishment. Peerless teacher, selflessly imparts expertise to subordinates, peers, and seniors.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

JUSTIFICATION:

E. INDIVIDUAL CHARACTER

1. COURAGE. Moral or physical strength to overcome danger, fear, difficulty or anxiety. Personal acceptance of responsibility and accountability, placing conscience over competing interests regardless of consequences. Conscious, overriding decision to risk bodily harm or death to accomplish the mission or save others. The will to persevere despite uncertainty.

ADV			N/O
Demonstrates inner strength and acceptance of responsibility commensurate with scope of duties and experience. Willing to face moral or physical challenges in pursuit of mission accomplishment.	Guided by conscience in all actions. Proven ability to overcome danger, fear, difficulty or anxiety. Exhibits bravery in the face of adversity and uncertainty. Not deterred by morally difficult situations or hazardous responsibilities.	Uncommon bravery and capacity to overcome obstacles and inspire others in the face of moral dilemma or life-threatening danger. Demonstrated under the most adverse conditions. Selfless. Always places conscience over competing interests regardless of physical or personal consequences.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

2. EFFECTIVENESS UNDER STRESS. Thinking, functioning and leading effectively under conditions of physical and/or mental pressure. Maintaining composure appropriate for the situation, while displaying steady purpose of action, enabling one to inspire others while continuing to lead under adverse conditions. Physical and emotional strength, resilience and endurance are elements.

ADV			N/O
Exhibits discipline and stability under pressure. Judgment and effective problem-solving skills are evident.	Consistently demonstrates maturity, mental agility and willpower during periods of adversity. Provides order to chaos through the application of intuition, problem-solving skills, and leadership. Composure reassures others.	Demonstrates seldom-matched presence of mind under the most demanding circumstances. Stabilizes any situation through the resolute and timely application of direction, focus and personal presence.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

3. INITIATIVE. Action in the absence of specific direction. Seeing what needs to be done and acting without prompting. The instinct to begin a task and follow through energetically on one's own accord. Being creative, proactive and decisive. Transforming opportunity into action.

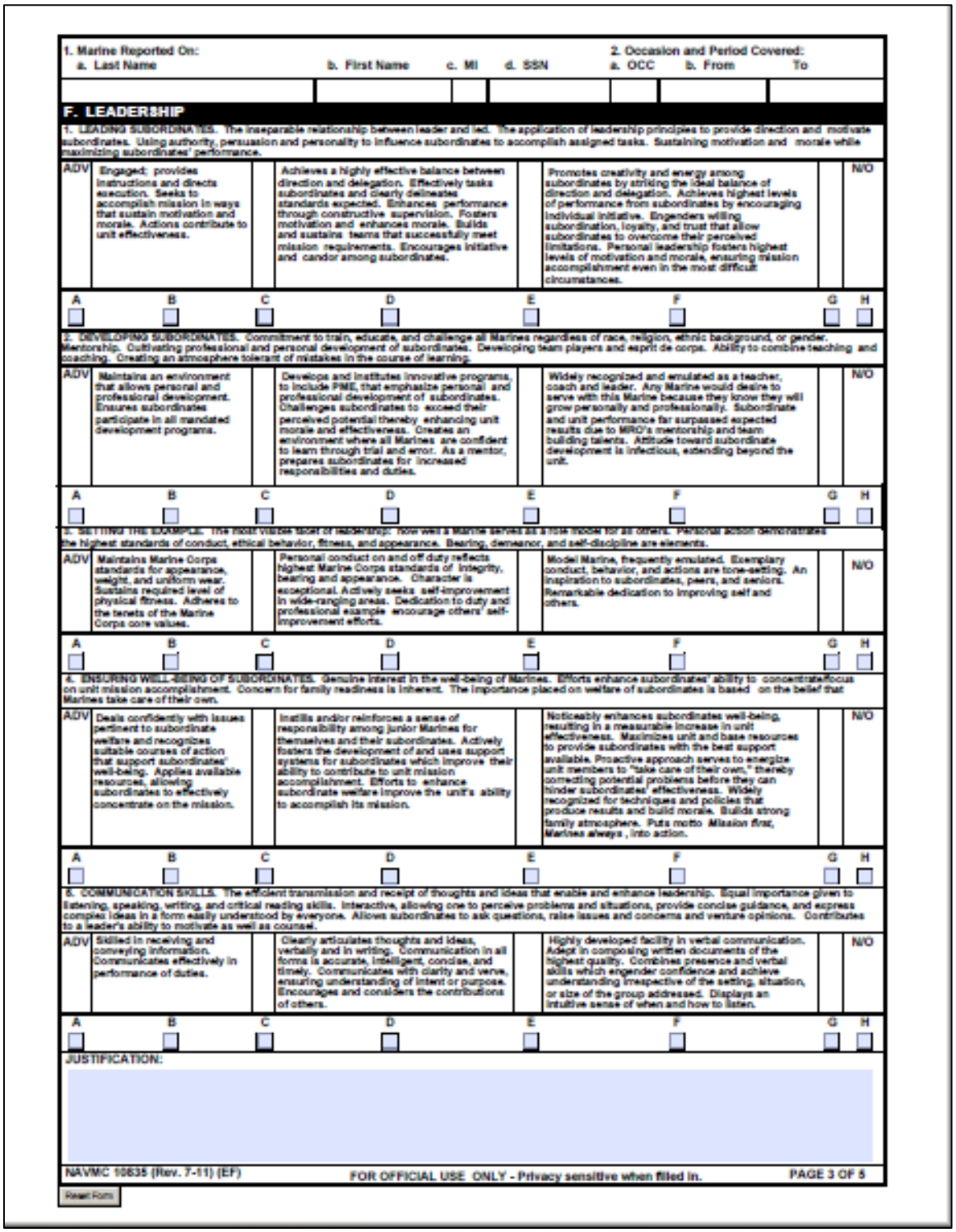
ADV			N/O
Demonstrates willingness to take action in the absence of specific direction. Acts commensurate with grade, training and experience.	Self-motivated and action-oriented. Foresight and energy consistently transform opportunity into action. Develops and pursues creative, innovative solutions. Acts without prompting. Self-starter.	Highly motivated and proactive. Displays exceptional awareness of surroundings and environment. Uncanny ability to anticipate mission requirements and quickly formulate original, far-reaching solutions. Always takes decisive, effective action.	

A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐

JUSTIFICATION:

NAVMC 10835 (Rev. 7-11) (EF) FOR OFFICIAL USE ONLY - Privacy sensitive when filled in. PAGE 2 OF 5

Reset Form



1. Marine Reported On: a. Last Name	b. First Name	c. MI	d. SSN	2. Occasion and Period Covered: a. OCC	b. From	To

F. LEADERSHIP

1. LEADING SUBORDINATES. The inseparable relationship between leader and led. The application of leadership principles to provide direction and motivate subordinates. Using authority, persuasion and personality to influence subordinates to accomplish assigned tasks. Sustaining motivation and morale while maximizing subordinates' performance.

ADV	Engaged; provides instructions and directs execution. Seeks to accomplish mission in ways that sustain motivation and morale. Actions contribute to unit effectiveness.		Achieves a highly effective balance between direction and delegation. Effectively tasks subordinates and clearly delineates standards expected. Enhances performance through constructive supervision. Fosters motivation and enhances morale. Builds and sustains teams that successfully meet mission requirements. Encourages initiative and candor among subordinates.		Promotes creativity and energy among subordinates by striking the ideal balance of direction and delegation. Achieves highest levels of performance from subordinates by encouraging individual initiative. Engenders willing subordination, loyalty, and trust that allow subordinates to overcome their perceived limitations. Personal leadership fosters highest levels of motivation and morale, ensuring mission accomplishment even in the most difficult circumstances.		N/O
A	B	C	D	E	F	G	H						
☐	☐	☐	☐	☐	☐	☐	☐						

2. DEVELOPING SUBORDINATES. Commitment to train, educate, and challenge all Marines regardless of race, religion, ethnic background, or gender. Mentorship. Cultivating professional and personal development of subordinates. Developing team players and esprit de corps. Ability to combine teaching and coaching. Creating an atmosphere tolerant of mistakes in the course of learning.

ADV	Maintains an environment that allows personal and professional development. Ensures subordinates participate in all mandated development programs.		Develops and institutes innovative programs, to include PME, that emphasize personal and professional development of subordinates. Challenges subordinates to exceed their perceived potential thereby enhancing unit morale and effectiveness. Creates an environment where all Marines are confident to learn through trial and error. As a mentor, prepares subordinates for increased responsibilities and duties.		Widely recognized and emulated as a teacher, coach and leader. Any Marine would desire to serve with this Marine because they know they will grow personally and professionally. Subordinate and unit performance far surpassed expected results due to MRO's mentorship and team building talents. Attitude toward subordinate development is infectious, extending beyond the unit.		N/O
A	B	C	D	E	F	G	H						
☐	☐	☐	☐	☐	☐	☐	☐						

3. SETTING THE EXAMPLE. The most visible facet of leadership: how well a Marine serves as a role model for all others. Personal action demonstrates the highest standards of conduct, ethical behavior, fitness, and appearance. Bearing, demeanor, and self-discipline are elements.

ADV	Maintains Marine Corps standards for appearance, weight, and uniform wear. Sustains required level of physical fitness. Adheres to the tenets of the Marine Corps core values.		Personal conduct on and off duty reflects highest Marine Corps standards of integrity, bearing and appearance. Character is exceptional. Actively seeks self-improvement in wide-ranging areas. Dedication to duty and professional example encourage others' self-improvement efforts.		Model Marine, frequently emulated. Exemplary conduct, behavior, and actions are tone-setting. An inspiration to subordinates, peers, and seniors. Remarkable dedication to improving self and others.		N/O
A	B	C	D	E	F	G	H						
☐	☐	☐	☐	☐	☐	☐	☐						

4. ENSURING WELL-BEING OF SUBORDINATES. Genuine interest in the well-being of Marines. Efforts enhance subordinates' ability to concentrate/focus on unit mission accomplishment. Concern for family readiness is inherent. The importance placed on welfare of subordinates is based on the belief that Marines take care of their own.

ADV	Deals confidently with issues pertinent to subordinate welfare and recognizes suitable courses of action that support subordinates' well-being. Applies available resources, allowing subordinates to effectively concentrate on the mission.		Instills and/or reinforces a sense of responsibility among junior Marines for themselves and their subordinates. Actively fosters the development of and uses support systems for subordinates which improve their ability to contribute to unit mission accomplishment. Efforts to enhance subordinate welfare improve the unit's ability to accomplish its mission.		Noticeably enhances subordinates well-being, resulting in a measurable increase in unit effectiveness. Maximizes unit and base resources to provide subordinates with the best support available. Proactive approach serves to energize unit members to "take care of their own," thereby correcting potential problems before they can hinder subordinates' effectiveness. Widely recognized for techniques and policies that produce results and build morale. Builds strong family atmosphere. Puts motto Mission first, Marines always , into action.		N/O
A	B	C	D	E	F	G	H						
☐	☐	☐	☐	☐	☐	☐	☐						

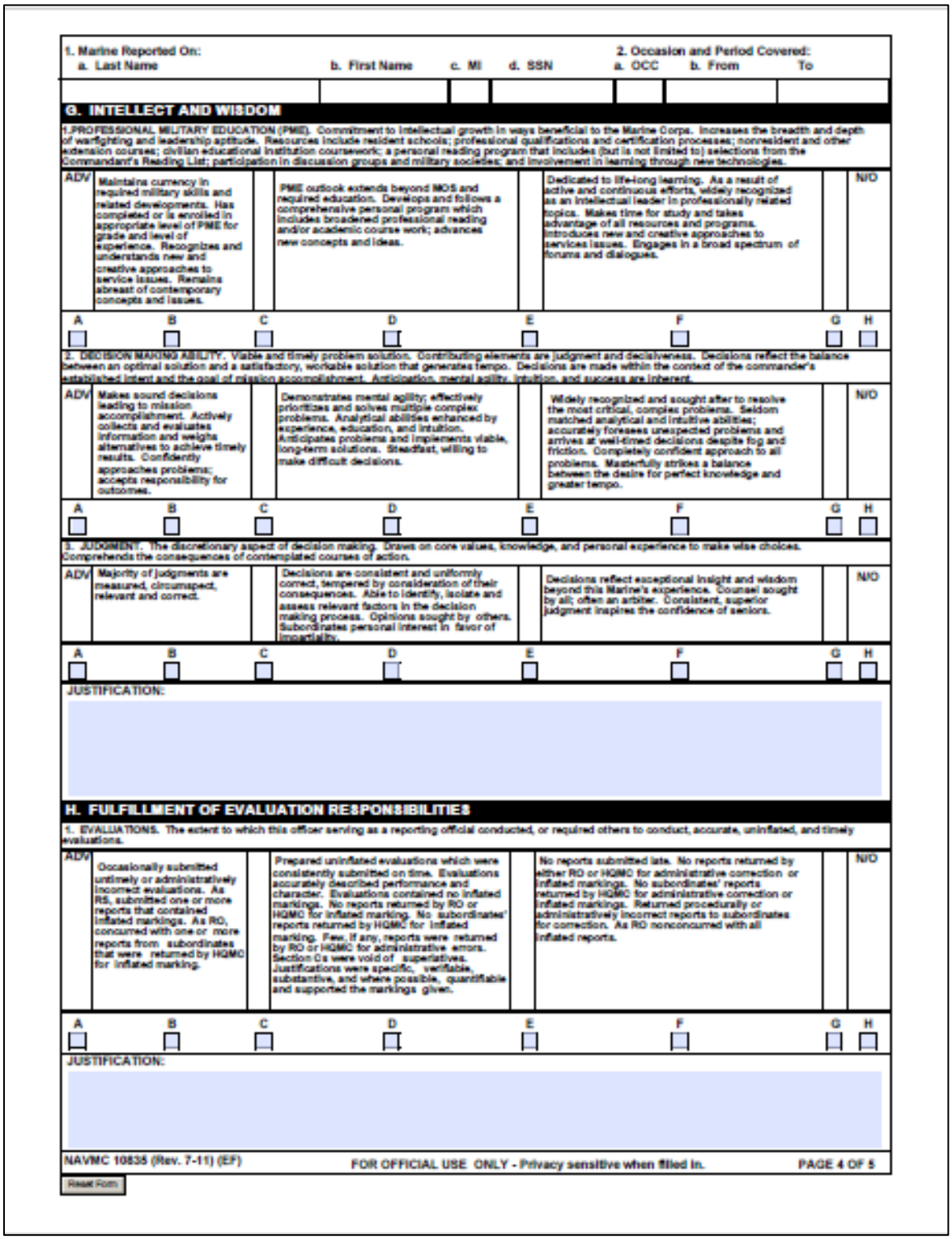
5. COMMUNICATION SKILLS. The efficient transmission and receipt of thoughts and ideas that enable and enhance leadership. Equal importance given to listening, speaking, writing, and critical reading skills. Interactive, allowing one to perceive problems and situations, provide concise guidance, and express complex ideas in a form easily understood by everyone. Allows subordinates to ask questions, raise issues and concerns and venture opinions. Contributes to a leader's ability to motivate as well as counsel.

ADV	Skilled in receiving and conveying information. Communicates effectively in performance of duties.		Clearly articulates thoughts and ideas, verbally and in writing. Communication in all forms is accurate, intelligent, concise, and timely. Communicates with clarity and verve, ensuring understanding of intent or purpose. Encourages and considers the contributions of others.		Highly developed facility in verbal communication. Adept in composing written documents of the highest quality. Combines presence and verbal skills which engender confidence and achieve understanding irrespective of the setting, situation, or size of the group addressed. Displays an intuitive sense of when and how to listen.		N/O
A	B	C	D	E	F	G	H						
☐	☐	☐	☐	☐	☐	☐	☐						

JUSTIFICATION:

NAVMC 10835 (Rev. 7-11) (EF) FOR OFFICIAL USE ONLY - Privacy sensitive when filled in. PAGE 3 OF 5

Reset Form



1. Marine Reported On: a. Last Name | b. First Name | c. MI | d. SSN | 2. Occasion and Period Covered: a. OCC | b. From | To

G. INTELLECT AND WISDOM

1.PROFESSIONAL MILITARY EDUCATION (PME). Commitment to intellectual growth in ways beneficial to the Marine Corps. Increases the breadth and depth of warfighting and leadership aptitude. Resources include resident schools; professional qualifications and certification processes; nonresident and other extension courses; civilian educational institution coursework; a personal reading program that includes (but is not limited to) selections from the Commandant's Reading List; participation in discussion groups and military societies; and involvement in learning through new technologies.

ADV									N/O
	Maintains currency in required military skills and related developments. Has completed or is enrolled in appropriate level of PME for grade and level of experience. Recognizes and understands new and creative approaches to service issues. Remains abreast of contemporary concepts and issues.		PME outlook extends beyond MOS and required education. Develops and follows a comprehensive personal program which includes broadened professional reading and/or academic course work; advances new concepts and ideas.		Dedicated to life-long learning. As a result of active and continuous efforts, widely recognized as an intellectual leader in professionally related topics. Makes time for study and takes advantage of all resources and programs. Introduces new and creative approaches to services issues. Engages in a broad spectrum of forums and dialogues.			
	A	B	C	D	E	F	G	H	

2. DECISION MAKING ABILITY. Viable and timely problem solution. Contributing elements are judgment and decisiveness. Decisions reflect the balance between an optimal solution and a satisfactory, workable solution that generates tempo. Decisions are made within the context of the commander's established intent and the goal of mission accomplishment. Anticipation, mental agility, intuition, and success are inherent.

ADV									N/O
	Makes sound decisions leading to mission accomplishment. Actively collects and evaluates information and weighs alternatives to achieve timely results. Confidently approaches problems; accepts responsibility for outcomes.		Demonstrates mental agility; effectively prioritizes and solves multiple complex problems. Analytical abilities enhanced by experience, education, and intuition. Anticipates problems and implements viable, long-term solutions. Steadfast, willing to make difficult decisions.		Widely recognized and sought after to resolve the most critical, complex problems. Seldom matched analytical and intuitive abilities; accurately foresees unexpected problems and arrives at well-timed decisions despite fog and friction. Completely confident approach to all problems. Masterfully strikes a balance between the desire for perfect knowledge and greater tempo.			
	A	B	C	D	E	F	G	H	

3. JUDGMENT. The discretionary aspect of decision making. Draws on core values, knowledge, and personal experience to make wise choices. Comprehends the consequences of contemplated courses of action.

ADV									N/O
	Majority of judgments are measured, circumspect, relevant and correct.		Decisions are consistent and uniformly correct, tempered by consideration of their consequences. Able to identify, isolate and assess relevant factors in the decision making process. Opinions sought by others. Subordinates personal interest in favor of impartiality.		Decisions reflect exceptional insight and wisdom beyond this Marine's experience. Counsel sought by all; often an arbiter. Consistent, superior judgment inspires the confidence of seniors.			
	A	B	C	D	E	F	G	H	

JUSTIFICATION:

H. FULFILLMENT OF EVALUATION RESPONSIBILITIES

1. EVALUATIONS. The extent to which this officer serving as a reporting official conducted, or required others to conduct, accurate, uninflated, and timely evaluations.

ADV									N/O
	Occasionally submitted untimely or administratively incorrect evaluations. As RS, submitted one or more reports that contained inflated markings. As RO, concurred with one or more reports from subordinates that were returned by HQMC for inflated marking.		Prepared uninflated evaluations which were consistently submitted on time. Evaluations accurately described performance and character. Evaluations contained no inflated markings. No reports returned by RO or HQMC for inflated marking. No subordinates' reports returned by HQMC for inflated marking. Few, if any, reports were returned by RO or HQMC for administrative errors. Section Cs were void of superlatives. Justifications were specific, verifiable, substantive, and where possible, quantifiable and supported the markings given.		No reports submitted late. No reports returned by either RO or HQMC for administrative correction or inflated markings. No subordinates' reports returned by HQMC for administrative correction or inflated markings. Returned procedurally or administratively incorrect reports to subordinates for correction. As RO nonconcurred with all inflated reports.			
	A	B	C	D	E	F	G	H	

JUSTIFICATION:

NAVMC 10835 (Rev. 7-11) (EF) FOR OFFICIAL USE ONLY - Privacy sensitive when filled in. PAGE 4 OF 5

Reset Form

1. Marine Reported On: a. Last Name | b. First Name | c. MI | d. SSN | 2. Occasion and Period Covered: a. OCC | b. From | To

I. DIRECTED AND ADDITIONAL COMMENTS

J. CERTIFICATION

1. I CERTIFY that to the best of my knowledge and belief all entries made hereon are true and without prejudice or partiality and that I have provided a signed copy of this report to the Marine Reported on.

(Signature of Reporting Senior) (Date in YYYYMMDD format)

2. I ACKNOWLEDGE the adverse nature of this report and

☐ I have no statement to make
☐ I have attached a statement

(Signature of Marine Reported On) (Date in YYYYMMDD format)

K. REVIEWING OFFICER COMMENTS

1. OBSERVATION: ☐ Sufficient ☐ Insufficient

2. EVALUATION: ☐ Concur ☐ Do Not Concur

3. COMPARATIVE ASSESSMENT: Provide a comparative assessment of potential by placing an "X" in the appropriate box. In marking the comparison, consider all Marines of this grade whose professional abilities are known to you personally.

DESCRIPTION	COMPARATIVE ASSESSMENT
THE EMINENTLY QUALIFIED MARINE	☐
ONE OF THE FEW	☐
EXCEPTIONALLY QUALIFIED MARINES	☐
ONE OF THE MANY HIGHLY QUALIFIED	☐
PROFESSIONALS WHO FORM THE MAJORITY OF THIS GRADE	☐
A QUALIFIED MARINE	☐
UNSATISFACTORY	☐

4. REVIEWING OFFICER COMMENTS: Amplify your comparative assessment mark; evaluate potential for continued professional development to include: promotion, command, assignment, resident PME, and retention; and put Reporting Senior marks and comments in perspective.

5. I CERTIFY that to the best of my knowledge and belief all entries made hereon are true and without prejudice or partiality.

(Signature of Reviewing Officer) (Date in YYYYMMDD format)

6. I ACKNOWLEDGE the adverse nature of this report and

☐ I have no statement to make
☐ I have attached a statement

(Signature of Marine Reported On) (Date in YYYYMMDD format)

L. ADDENDUM PAGE

ADDENDUM PAGE ATTACHED: ☐ YES

NAVMC 10835 (Rev. 7-11) (EF) FOR OFFICIAL USE ONLY - Privacy sensitive when filled in. PAGE 5 OF 5

Reset Form

APPENDIX C. INTER-RATER RELIABILITY SURVEY

Intro

USMC Fitness Report Inter-Rater Reliability

Welcome to the FitRep Content Reliability Survey. Your insights are invaluable in contributing to a comprehensive understanding of the United States Marine Corps' Performance Evaluation System. This research focuses on the reliability of the Fitness Report (FitRep) content, examining the inter-rater perspective among Reporting Seniors (RSs) like yourself.

Informed Consent

INFORMED CONSENT

Privacy Notice: Your privacy and voluntary participation in this survey are of utmost importance. We want to assure you that all information collected in this survey will be treated with strict confidentiality. Your Personal Identifiable Information, including names and contact details, will not

be maintained or linked to your survey responses.

Voluntary Participation: Your participation in this survey is entirely voluntary. You have the right to discontinue your participation at any time without facing any consequences. By selecting "Begin Survey" and continuing with this survey, you are providing informed consent to participate. Please be aware that your decision to participate, or not, will not impact your relationship with any organization or institution associated with this research.

Data Usage: The data collected through this survey will be used solely for research purposes related to assessing the inter-rater reliability of Marine Corps Fitness Reports. Rest assured that your information will be handled with the utmost care and only used for this specific research project. If you have any questions or concerns about this survey or your participation, please do not hesitate to contact the student investigator at fatima.banks@nps.edu.

Background Information

BACKGROUND INFORMATION

The following questions contain information about your socio demographic information which is being collected to infer correlation between officer characteristics and responses. There will be no attempt to identify individual respondents based on the information provided.

Please identify your rank.

2ndLt 1stLt Capt Maj Col LtCol

Please identify your 4-digit MOS (i.e. 3404).

Select your appropriate MOS category

- Combat Arms
- Combat Service Support
- Aviation
- Aviation Ground Support
- Legal
- Intelligence

○ Special Forces/Operations
○ Other

Please indicate your level of completed training (formal or informal) on the subject of performance evaluations.

	TBS Entry-Level Training	Command-Level Training	Other Training
Choose One	○	○	○

Please indicate the number of individual Marines you have evaluated at each grade (number can be an estimate).

	0	1-5	6-10	11+
E-5	○	○	○	○
E-6 - E-7	○	○	○	○
E-8 - E-9	○	○	○	○
O-1 - O-3	○	○	○	○
O-4 - O5	○	○	○	○

Scenario

Scenario
Put yourself in the role of the Regiment Executive Officer for Regiment X. You have been on deck for 6 months. You are the Reporting Senior for "Marine Officer" the Regimental Supply Officer (SuppO). Marine Officer has submitted their annual Marine Reported-On Worksheet (MROW) for evaluation, and your responsibility is to perform a thorough and unbiased assessment based on your personal grading philosophy. Your task involves accurately and fairly rating eight attributes using the Performance-Anchored Rating Scales (PARS).

Background:
"Ground Supply Officers are a special Staff Officer that supervises the Commanders' Property, Plant, & Equipment (PP&E) and Operating Material & Supplies (OM&S) to ensure data accuracy, existence, and completeness (equipment accountability, visibility, and auditability). They supervise and coordinate ground supply administration and operations for supply activities, units, bases, or stations, to include operating forces and shore station organizations. Ground Supply Officers may direct the activities of a maintenance distribution or industrial type organization. They command or serve in either an operating forces service unit or a non-operating forces activity. Ground Supply Officers supervise the execution of supply chain

management policies and procedures pertaining to: procurement; receipt; inventory control; repair; storage; distribution; issue; disposal; and computation and maintenance of stock positioning requirements. They provide supply support insight for operational planning requirements; supervise transportation of supplies and equipment; manage the transmittal of public funds; participate in the budget process, administer, and expend allotted funds; and make necessary recommendations to the Commanding Officer regarding supply support procedures." (Marine Officer MOS Assignment Handbook, 2019, p81)

To initiate the evaluation, access the MROW by selecting the attachment labeled "MROW_Marine Officer." You will review the provided information to complete the assessment. For your reference, pages 2-6 of the attachment labeled "Fitness Report_Marine Officer, contain descriptions of the attributes and PARS. Please note that all responses should be submitted through the survey, and refrain from rating attributes on the attached NAVMC 10835.

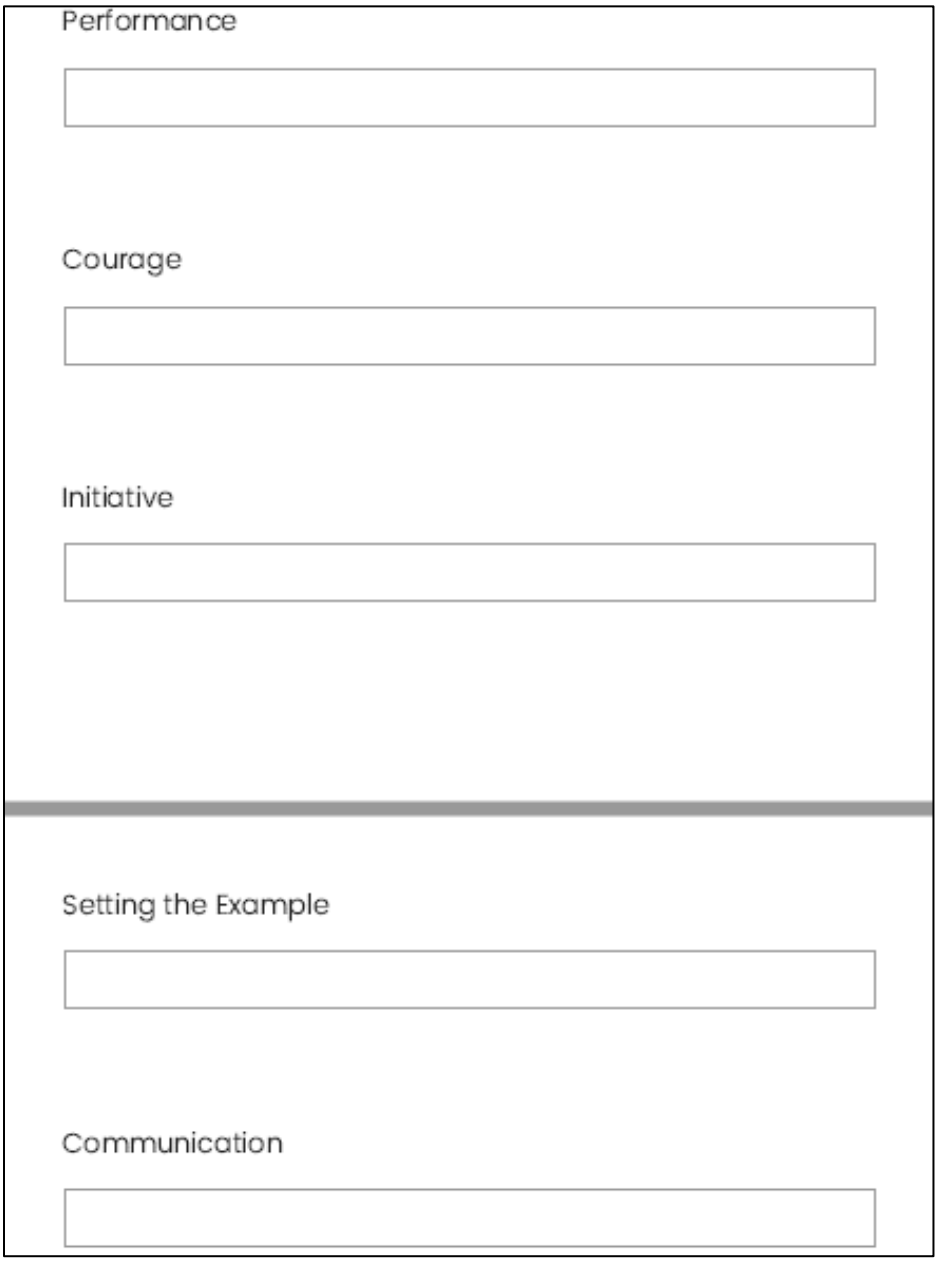
Using the attribute and rating descriptions outlined in sections D through G of the NAVMC 10835 (pages 2-4), rate the accomplishments of the MRO as identified in sections A through C, with the appropriate letter score.

	A	B	C	D	E	F	G	H
Performance	○	○	○	○	○	○	○	○
Courage	○	○	○	○	○	○	○	○
Initiative	○	○	○	○	○	○	○	○
Setting the Example	○	○	○	○	○	○	○	○
Communication	○	○	○	○	○	○	○	○
PME	○	○	○	○	○	○	○	○
Decision-Making Ability	○	○	○	○	○	○	○	○
Judgement	○	○	○	○	○	○	○	○

Outline the particular aspects of the MRO's accomplishments in the MROW that influenced your rating. Provide one or two sentences illustrating how the MRO's

achievements significantly impacted your assessment of their performance in this specific personal attribute.

In instances where accomplishments fall short of supporting a comprehensive evaluation, specify the behaviors you believe reflect this particular attribute.


Performance
Courage
Initiative
Setting the Example
Communication

APPENDIX D. SURVEY EMAIL INVITATION

Invitation to Participate in Thesis Survey on USMC Performance Evaluation System

Banks, Fatima (Capt) <fatima.banks@nps.edu>
Wed 1/31/2024 10:25 AM

To: **NPS USMC <NPSUSMC@NPS01.onmicrosoft.com>**
Cc: Seagren, Chad (CIV) <cwseagre@nps.edu>

Dear Fellow Marines,

I am reaching out to invite you to participate in a research initiative that aims to enhance the understanding of the United States Marine Corps' Performance Evaluation System (PES).

Survey Details
Title: FitRep Content Reliability: An Inter-Rater Perspective
Purpose: Assess the reliability of FitRep content, focusing on the inter-rater perspective among Reporting Seniors.
Duration: Approximately 15-20 minutes (19 questions)
Empirical research on the Performance Evaluation System (PES) has uncovered unexplained variations in Reporting Senior (RS) scoring, raising concerns about the fairness and objectivity of the fitness report (FitRep). By participating in this survey, you play a key role in identifying potential causal factors that may influence the statistical reliability of FitRep attribute ratings—absent the RS profile. Survey results allow us to comprehensively test and enhance the assessment process. The survey aims to identify the consistency and dependability of Personal Anchored Rating Scale (PARS) ratings for individual attributes assessed on the FitRep. Your unique insights as Marine Corps Officers, holding the inherent responsibility of a Reporting Senior, are invaluable to this study. Your participation is crucial to ongoing efforts aimed at improving the organizational effectiveness and fairness of the evaluation process.

How to Participate:
Please follow the below link to access the survey. Your responses will be strictly confidential, and your candid feedback is highly appreciated.

Survey Link: https://navalpostgradfedramp.gov1.qualtrics.com/jfe/form/SV_82nSOkBk1fYVvv0

We kindly request your participation by 25 February 2024. Thank you in advance for your time and consideration. Your participation is a vital contribution to the success of this research initiative. Should you have any questions, concerns, or comments please feel free to reach out.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Balakrishnan, R., Lin, H. & Sivaramakrishnan, K. (2016). *On the relative efficacies of ranking and absolute performance evaluation systems*. SSRN. https://ssrn.com/abstract=2828956 or http://dx.doi.org/10.2139/ssrn.2828956

Ball, H. L. (2019). Conducting online surveys. *Journal of Human Lactation, 35*(3), 413–417. https://doi.org/10.1177/0890334419848734

Berger, D. H. (2019). *Commandant's Planning Guidance: 38th Commandant of the Marine Corps*. https://www.hqmc.marines.mil/Portals/142/Docs/%2038th%20Commandant's%20Planning%20Guidance_2019.pdf

Bottoms, L. (2022). *A systematic approach to interrater reliability in early childhood teacher performance evaluations*. [Doctoral dissertation, University of North Carolina at Charlotte]. https://www.proquest.com/docview/2660956629?parentSessionId=XAcQfOawOGoTdtFX9PdgKD%2BlfRVGD7qqfo8mSO4g6Kg%3D&sourcetype=Dissertations%20&%20Theses

Chachula, S. K. (1992). *Performance measurement systems: A best practices study*. [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/24005

Clemens, A., Malone, L., Phillips, S., Lee, G., Hiatt, C. & Kimble, T. (2012). *An Evaluation of the Fitness Report System for Marine Officers*. (Report No. DRM-2012-U-001003-Final). Center for Naval Analyses. https://www.hqmc.marines.mil/Portals/138/DRM-2012-U-001003-Final.pdf

Dunst, W. L. (2018). *Evolution of the Marine officer fitness report: a multivariate analysis*. [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/5829

Euske, K. J., Lebas, M. J., & McNair, C. J. (1993). Performance management in an international setting. *Management Accounting Research, 4*(1), 3-20. https://doi.org/10.1006/mare.1993.1016

Gómez-Mejia, L. R., Balkin, D. B. & Cardy, R. L. (2016). Managing human resources. Harlow: Pearson.

Gupta, S. (2022). *Anscombe's Quartet: What is it and why do we care?* Built In. https://builtin.com/data-science/anscombes-quartet

Headquarters, United States Marine Corps (2023). *Performance Evaluation System (PES)* (MCO 1610.7B). https://www.marines.mil/Portals/1/Publications/MCO%201610.7B%20SECURED.pdf?ver=Y104Ok-51cS4PJssb_4t4g%3d%3d

Heuer, J. W. (2020). The Marine Corps promotion board process: An after-action report from a board member. *Marine Corps Gazette*.https://www.mca-marines.org/wp-content/uploads/49-The-Marine-Corps-Promotion-Board-Process.pdf

Jobst, M. G., & Palmer, J. (2005). *An analysis of the USMC FITREP: Contemporary or inflexible?* [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle.net/10945/2210

Koo, T.K. & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med,15*(2):155-63. doi: 10.1016/j.jcm.2016.02.012.

Larger, R. B., Jr. (2017). *Effectiveness of the Marine Corps' junior enlisted performance evaluation system: An evaluation of proficiency and conduct marks* [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/53006

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLoS One*. *22*;14(7). doi: 10.1371/journal.pone.0219854

Luke, B. C. (2022). *Performance evaluation trait validation*. [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/69677

Mitani, A. A, Freer, P. E. & Nelson, K. P. (2017). Summary measures of agreement and association between many raters' ordinal classifications. *Ann Epidemiol*, *27*(10), 677–685. doi: 10.1016/j.annepidem.2017.09.001.

National Research Council. (1991). *Performance assessment for the workplace: Volume I*. Washington, DC: The National Academies Press. https://nap.nationalacademies.org/read/1862/chapter/8?term=reliability

Nichols, T. R., Wisner, P. M., Cripe, G. & Gulabchand, L. (2010). Putting the Kappa statistic to use. *Quality Assurance Journal*,*13*, 57–61. https://doi.org/10.1002/qaj.481

Phillips, S., & Clemens, A. (2011). *The Fitness Report System for Marine Officers: Prior Research* (Report No. CIM D0026273.A1/Final). Center for Naval Analyses. https://www.cna.org/reports/2011/D0026273.A1.pdf

Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher, 27*(2). https://mwera.org/MWER/volumes/v27/issue2/v27n2-Pufpaff-FEATURE-ARTICLE.pdf

Rigaut, P. (2017). *A text analysis of the Marine Corps fitness report*. [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://apps.dtic.mil/dtic/tr/fulltext/u2/1046515.pdf

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs. relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*(3), 302–316. https://doi.org/10.1111/j.1468-2389.2007.00390.x

Shrotryia, V. K., & Dhanda, U. (2019). Content validity of assessment instrument for employee engagement. *SAGE Open*, 9(1). https://doi.org/10.1177/2158244018821751

Siegrist, K. (2022). *6.3*: *The law of large numbers.* LibreOne Launchpad. https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/06%3A_Random_Samples/6.03%3A_The_Law_of_Large_Numbers

Small, L. C. (2020). *Successful practices for employee performance evaluations*. [Master's Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/64880

Sreedhara, V. S. (2015). Control of Thermoforming Process Parameters to Increase Quality of Surfaces Using Pin-Based Tooling. *Proceedings of the 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference.* https://doi.org/10.1115/DETC2015-47682

Stiles, P., & Kulvisaechana, S. (n.d). *Human capital and performance: A literature review*. http://www.bus.tu.ac.th/usr/sab/articles_pdf/research_papers/dti_paper_web.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Fort Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California