2017-03

Effectiveness of the Marine Corps' junior enlisted performance evaluation system: an evaluation of proficiency and conduct marks

Larger, Richard B., Jr.

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

EFFECTIVENESS OF THE MARINE CORPS' JUNIOR ENLISTED PERFORMANCE EVALUATION SYSTEM: AN EVALUATION OF PROFICIENCY AND CONDUCT MARKS

by

Richard B. Larger Jr.

March 2017

Thesis Advisor: Chad Seagren
Co-Advisor: Marigee Bacolod

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704–0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2017	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
EFFECTIVENESS OF THE MARINE CORPS' JUNIOR ENLISTED PERFORMANCE EVALUATION SYSTEM: AN EVALUATION OF PROFICIENCY AND CONDUCT MARKS

5. FUNDING NUMBERS

6. AUTHOR(S) Richard B. Larger Jr.

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)
Manpower & Reserve Affairs
Quantico, VA 22134-5103

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
NPS-17-M001-A

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. IRB number **NPS.2017.0013-IR-EP5-A.**

12a. DISTRIBUTION / AVAILABILITY STATEMENT	**12b. DISTRIBUTION CODE**
Approved for public release. Distribution is unlimited.	

13. ABSTRACT (maximum 200 words)

This thesis analyzes the effectiveness of the U.S. Marine Corps' proficiency and conduct marks as measures of job performance for promotion decisions. The analysis uses big data techniques (factor analysis) and multivariate regressions on data of 360,690 active duty Marines who held the paygrade of E3 or E4 between 2006 and 2016 to estimate the reliability, validity, accuracy, and practicality of proficiency and conduct marks.

Overall, results show that proficiency and conduct marks are effective indicators of performance, with some room for improvement. Marks are statistically inconsistent between raters, and proficiency and conduct marks essentially measure the same type of performance. The factor analysis does show that proficiency and conduct marks together are the most important factors in the composite score for E4s and the second most important, behind experience, for E3s. Lastly, proficiency and conduct marks are the most predictive of future performance compared to all other composite score variables.

The author recommends that the Marine Corps continue to use proficiency and conduct marks as a basis for promotion decisions, but that the Marine Corps should redefine the marks in order to improve interpretability and minimize redundancies.

14. SUBJECT TERMS performance evaluation, proficiency marks, conduct marks, Marine Corps enlisted performance evaluation system	**15. NUMBER OF PAGES** 117
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

EFFECTIVENESS OF THE MARINE CORPS' JUNIOR ENLISTED PERFORMANCE EVALUATION SYSTEM: AN EVALUATION OF PROFICIENCY AND CONDUCT MARKS

Richard B. Larger Jr.
Captain, United States Marine Corps
B.B.A., American Military University, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2017

Approved by: Chad Seagren
Thesis Advisor

Marigee Bacolod
Co-Advisor

William Hatch
Academic Associate
Graduate School of Business and Public Policy

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This thesis analyzes the effectiveness of the U.S. Marine Corps' proficiency and conduct marks as measures of job performance for promotion decisions. The analysis uses big data techniques (factor analysis) and multivariate regressions on data of 360,690 active duty Marines who held the paygrade of E3 or E4 between 2006 and 2016 to estimate the reliability, validity, accuracy, and practicality of proficiency and conduct marks.

Overall, results show that proficiency and conduct marks are effective indicators of performance, with some room for improvement. Marks are statistically inconsistent between raters, and proficiency and conduct marks essentially measure the same type of performance. The factor analysis does show that proficiency and conduct marks together are the most important factors in the composite score for E4s and the second most important, behind experience, for E3s. Lastly, proficiency and conduct marks are the most predictive of future performance compared to all other composite score variables.

The author recommends that the Marine Corps continue to use proficiency and conduct marks as a basis for promotion decisions, but that the Marine Corps should redefine the marks in order to improve interpretability and minimize redundancies.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AAV	amphibious assault vehicle
BARS	behaviorally anchored rating scale
BCRM	basic combat rifle marksmanship
C2	command and control
CFT	combat fitness test
CLEP	College Level Examination Program
DI	drill instructor
FOR	frame of reference
FRM	fundamental rifle marksmanship
GMP	general military proficiency
HQMC	Headquarters Marine Corps
IRAM	*Individual Records Administrative Manual*
MCI	Marine Corps Institute
MCSF	Marine Corps Security Forces
MCU	Marine Corps University
MI	Manpower Information Systems Division
MMRP	Manpower Management Records and Performance branch
MSG	Marine security guard
NCO	noncommissioned officer
OccFld	occupational field
PFT	physical fitness test
PMOS	primary military occupational specialty
RSRV	reporting senior relative value
ROCV	reviewing officer cumulative value
TFDW	Total Force Data Warehouse
TIG	time in grade
TIS	time in service
USMC	Unites States Marine Corps

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

Foremost, I want to acknowledge how truly blessed I am. I have a Lord with unfailing love, a wife who is fully devoted to loving and supporting me, and two daughters who are incredibly resilient. I attribute my success to each of them.

My advisors, Drs. Chad Seagren and Marigee Bacolod, have my utmost gratitude for their dedication, mentorship, and support throughout the entire thesis process. I also am grateful for Tim Johnson, whose patience and timely support led to the bulk of the data for this research. I also extend my appreciation to Captains Emilie Monaghan, Will Wathen, and Greg Moynihan at Manpower Studies and Analysis Branch for chasing down all of my obscure requests and putting me in touch with all the right people. Finally, I want to thank Sergeant Major VanOostrom for his invaluable input and inspiration.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The purpose of this research is to determine if proficiency and conduct marks are *effective* measures of performance that lead to fair evaluations and promotions of junior enlisted Marines based on their true performance. The most recent related study, conducted in 1996 by the Marine Corps' Manpower & Reserve Affairs Department, addressed perceived issues of inflation and found that proficiency and conduct are relevant indicators of performance (W. Wathen, personal communication, October 25, 2016).

The proficiency and conduct mark system applies to all E4s and below in the Marine Corps and are the only subjective measures of performance within the evaluation system. All other performance is measured by rifle marksmanship score, physical fitness scores in the physical fitness test and combat fitness test, self-education, the Marine's special duty assignment status, and time spent in the Marine Corps as well as in current grade. Proficiency and conduct marks, together with the quantitative measures just listed, form the Marine's composite score for promotion. Proficiency and conduct marks are an input for other administrative decisions as well, such as retention, selection to competitive programs, and characterization of service upon discharge.

A. RESEARCH QUESTIONS BASED ON ACADEMIC LITERATURE

Literature reveals that the objective of any performance evaluation system should be to translate, as closely as possible, true work performance into a performance evaluation score. The literature also provides information on what affects the translation of performance into scores; this study focuses on the performance measure itself, the rating format, and the rater.

Based on a review of academic literature, this study defines effectiveness in terms of reliability, validity, accuracy, and practicality. The first three measures relate directly to the informational quality of the marks, and the latter opens the discussion to how usable and interpretable the marks are for making promotion decisions. This study addresses the following research questions based on academic literature. The first two

research questions pertain to reliability, the next two pertain to validity, and the final three pertain to accuracy.

1. (R1) Are proficiency and conduct marks consistent measures of performance over time?
2. (R2) Do proficiency and conduct marks vary between rater?
3. (V1) Which composite score variables provide the most information on the Marine's performance level?
4. (V2) Do proficiency and conduct marks predict future performance as indicated by fitness report scores?
5. (A1) Do proficiency and conduct marks differentiate between levels of performance?
6. (A2) Are proficiency and conduct marks subject to rater leniency?
7. (A3) Are proficiency and conduct marks distinct measures of performance?

To complete the analysis, this study uses semi-annual snapshots of demographic, performance, and occupational data for every active duty Marine at the paygrade of E1 to E4 from 2006 to 2016. Total Force Data Warehouse provided those data. In addition, to analyze predictive validity, this study uses fitness report data from Manpower Management Records and Performance Branch (MMRP) on all active duty Marines who promoted to sergeant between 2010 and 2013.

B. FINDINGS ON THE EFFECTIVENESS OF PROFICIENCY AND CONDUCT MARKS

Results show that proficiency and conduct marks are marginally effective performance measures. That is, they seem to be working as intended for the most part but need improvement to ensure fair promotions. Table 1 lists the literature-based hypotheses and results. Most noteworthy are the validity results for research questions V1 and V2. Proficiency and conduct marks are the strongest predictors of future performance compared to other performance measures, and they are the most important performance measures in the composite score for promotion-eligible E4s.

Table 1. Hypotheses on the Effectiveness of Proficiency and Conduct Marks.

Research Question	Null Hypotheses—Reliability, Validity, and Accuracy	Literature Suggests…	Supporting Evidence?
R1	Pro/con marks are stable	Yes	Yes
R2	Pro/con marks are consistent between raters	No	Inconclusive
V1	Pro/con marks are important contributions to a Marine's composite score	Yes	Yes
V2	Pro/con marks predict future performance	Yes	Yes
A1	Pro/con marks differentiate between levels of performance	No	Yes
A2	Pro/con marks are not inflated	No	No
A3	Pro/con marks are distinct measures of performance	No	No

1. Reliability

This study uses descriptive statistics to analyze consistency over time, or stability. Specifically, the standard deviation trends for both marks provide information on the marks' stability. Proficiency and conduct marks appear to be stable, though data limitations prevent this study from presenting evidence that is more concrete.

Multivariate regression analysis explains the effect of unit assignment on proficiency and conduct marks in order to reveal if different grading philosophies significantly affect the expected value of proficiency and conduct marks. For four groups of Marines separated by military occupational specialty, the analysis holds constant demographic, performance, and occupational variables to determine the effect. There is evidence that different unit types (e.g., ground and aviation) have different grading philosophies. The effect is small, however, and may not significantly affect promotion timing between two equally performing Marines. Nonetheless, further analysis is necessary to ascertain that promotions are not being affected.

2. Validity

This study uses factor analysis to explore the construct of the composite score and to determine which variables help explain a Marine's overall performance as a lance corporal or corporal. The latent performance variables revealed by factor analysis are

included in the model for predictive validity in order to analyze the latent variables' effect on the Marine's fitness report score in up to their first three years as a sergeant. The factor containing proficiency and conduct marks is the strongest predictor of future performance in terms of fitness report scores. In addition, of all composite score variables, proficiency and conduct marks provide the most information on a Marine's performance level in the sample containing promotion-eligible corporals.

3. Accuracy

This study uses the distribution of average marks and univariate regression results to estimate accuracy. The only result contrary to the hypotheses is that proficiency and conduct marks are differentiating between levels of performance. Proficiency and conduct marks are inflated but most likely to the benefit of differentiating between performance levels. Additionally, results show that proficiency and conduct are measuring much of the same performance, which is of concern if the marks intend to measure different aspects of a Marine's performance. Thus, further analysis is required to determine the relevancy of proficiency and conduct marks in terms of what they are intended to measure.

4. Practicality

This study provides an assessment of practicality based on the results as they pertain to the marks' interpretability, observability, and usability. Results suggest that raters have difficulty interpreting the marks, indicated by low interrater reliability, and that both marks are measuring much of the same performance. Regarding the marks' observability, predictive validity tests show that raters are able to observe the performance associated with the marks' definitions and multiple traits. The observed behaviors are similar to the performance attributes in a fitness report. Lastly, the results of this study do not provide a clear answer on the marks' usability for the purpose of promotion. They seem to do a good job identifying the Marines who have the most potential to perform as a sergeant, yet the inconsistency in grading philosophies may lead to unfair promotions. Further analysis is required to better assess the marks' usability. Table 2 summarizes the assessment of practicality.

Table 2. Hypotheses on the Practicality of Proficiency and Conduct Marks.

Hypotheses—Practicality	Supporting Hypotheses	Literature Suggests…	Supporting Evidence?
Raters can easily interpret the rating format	R2, A3	No	No
Raters are able to infer traits from observed behavior	V2	No	Yes
Pro/con marks are usable for the purpose of promotion decisions	R2, V1, V2, A1	Inconclusive	Inconclusive

C. RECOMMENDATIONS TO IMPROVE THE EFFECTIVENESS OF PROFICIENCY AND CONDUCT MARKS

This study provides recommendations based on a broad review of the performance evaluation system. Foremost, this study recommends further analysis to determine if the magnitude of the effects is justification to make significant changes. The first recommendation is the most actionable based on the results of this study. The other two require further supporting analysis.

1. Keep Proficiency and Conduct Marks and Improve the Interpretability of the Rating Format

There is value in having subjective performance measures in the evaluation process. Subjective measures are necessary to standardize evaluations in many different work environments across the Marine Corps. The rating format needs improvement by better defining the performance characteristics applicable to proficiency and conduct. The current system fails to differentiate between the two. If stakeholders choose to redefine the marks, they should avoid the use of traits (e.g., intellect and wisdom, adaptability) that are more difficult to observe than behaviors or results. Additionally, the rating format might also allow supervisors to assign scores to each of the attributes that pertain to proficiency and to each of the attributes that pertain to conduct. Those scores would combine to form an overall proficiency rating and an overall conduct rating.

Another way to improve interpretability is to redefine proficiency and conduct. Currently, both marks measure much of the same performance. First, subject matter experts should agree on the intent of both marks. Potentially, one could redefine

proficiency to focus mostly on the Marine's progression within primary military occupational specialty (PMOS) or primary duty. One could redefine conduct to focus mostly on the Marine's behavior unrelated to his or her specialty or primary duty.

2. Expand on Training Given at Professional Military Education Courses to Include Education Related to Cognitive Biases

In addition to improving the rating format, training supervisors on how to avoid cognitive-related biases can also improve rating accuracy (Pursell, Dossett, & Latham, 1980). The Marine Corps University could incorporate this type of training into existing Professional Military Education (PME) courses. Training could also focus on the proper use of the rating scale in scenario-based training to avoid inflation and inconsistencies between evaluators. However, designing effective curricula requires subject matter experts to explore the specific training needs. In addition, subject matter experts first need to estimate the value of said training and determine if the costs associated with training are justified.

3. Move Proficiency and Conduct Marks into Marine Corps Order 1610.7, *Performance Evaluation System*, Instead of the *Marine Corps Individual Records Administration Manual*.

Moving proficiency and conduct marks from Manpower Information Systems Division (MI), who authors and manages the *Marine Corps Individual Records Administration Manual* (IRAM), to MMRP, who authors and manages the fitness report system, will

> further allow the Marine Corps the ability to professionalize a performance evaluation continuum by combining performance evaluations under one Branch within the Manpower Management Division. However, additional costs related to expanding the role of MMRP will need to be considered such as systems integration and increases in the manpower workforce for MMRP. (R. VanOostrom, personal communication, February 17, 2017)

II. BACKGROUND

The purpose of the Marine Corps' junior enlisted performance evaluation system is to provide information about who to promote and retain (United States Marine Corps [USMC], 2010, 2012). The information is intended to represent actual performance behaviors and results that affect organizational goals. In this system, the information is gleaned from a combination of objective and subjective performance measures. This study focuses on the subjective performance measures—duty proficiency and conduct marks—though it briefly introduces the accompanying quantitative (objective) measures that affect promotion outcomes. All factors are considered together for analysis in a later chapter.

The *Marine Corps Individual Records Administration Manual (Short Title: IRAM)* (USMC, 2000) is the governing document for Marine Corps records and administration, including the assignment of proficiency and conduct marks. MI distributes and manages the IRAM. MI is responsible for managing functions pertaining to personnel administration (https://www.manpower. usmc.mil). It is worth mentioning that the IRAM is a peculiar document for proficiency and conduct marks to reside in, especially considering that the Marine Corps' other performance evaluation system pertaining to all other Marines is managed by the MMRP.

A. PROFICIENCY AND CONDUCT MARKS

Proficiency and conduct marks are a subjective assessment of the Marine's performance during a specified period. Marines are assigned proficiency and conduct marks from the rank of private through corporal. The marks are used to determine promotion score, retention competitiveness, characterization of service upon discharge, and eligibility for some special duty assignments and enlisted to officer commissioning selection.

According to the IRAM (USMC, 2000), a Marine is required to receive proficiency and conduct marks for the following reporting occasions:

- When ending a semiannual period for active duty (1 January and 1 July) or ending an annual period for reserve
- When transferred (e.g., reserve to active duty, active duty to reserve, completion of recruit training and initial skill training)
- When assigned to the Temporary Disability Retired List
- When discharged
- When promoted to corporal (E-4) or sergeant (E-5)
- When reduced in grade
- When declared a deserter (first day of unauthorized absence and then again on last day prior to declaring deserter)
- When assigned to temporary duty, one occasion prior to transfer and one upon completion
- When primary duty has changed
- When service school is completed
- When recommended per enlisted promotion manual (USMC, 2012)

In some instances, a Marine may meet two or more criteria for receiving new marks within a short period of time. The highest precedence occasion, outlined in the IRAM, is reported. All other lower precedent reporting occasions within 90 days following the previously assigned marks receive "NA" marks (USMC, 2000). A Marine receives marks at least every six months (semiannual reporting occasion) unless the Marine has a higher precedent reporting occasion within the past 90 days of the semiannual reporting occasion. Thus, a Marine receives marks at least every 270 days.

Commanders are responsible for assigning marks, which are based on recommendations from the Marine's more immediate supervisors (USMC, 2000), as shown in Figure 1. Depending on unit standard operating procedure, the Marine's immediate supervisor (also known as the Marine's noncommissioned officer [NCO]) typically generates marks, which are routed through the chain of command to the commanding officer via Marine Online or other means. Recommended marks may be altered at any step prior to approval to ensure the Marine is receiving marks

commensurate with their performance on and off duty and also to correct for inflation. Recommended marks may also be accompanied by comments to provide additional justification to the commander.

Noncommissioned Officer	Commander
Trains and supervises Marines in the performance of their duties.	Sets unit policies and procedures per the IRAM.
Records and evaluates Marines during the reporting period.	Consult with SNCOs/NCOs during the assignment process.
Provides recommendations to the commander.	Assign proficiency and conduct marks to Marines.

Figure 1. NCO and Commander Responsibilities.
Source: Marine Corps University (MCU; 2012).

1. Duty Proficiency Marks

Duty proficiency marks measure the Marine's performance in their primary duties. Marks range from 0.0 to 5.0. Figure 2 describes the duty proficiency marks' corresponding adjective ratings and narratives. According to the IRAM (USMC, 2000), the following is what should be considered when assigning duty proficiency marks:

> In addition to technical skills and specialized knowledge, relating to duty proficiency marks, the "whole Marine concept" must be considered. Such attributes as mission accomplishment, leadership, intellect and wisdom, individual character, physical fitness, personal appearance, and completion of professional military education, Marine Corps Institute courses, and off duty education should also be evaluated and incorporated into the duty proficiency mark. (p. 4–42)

Additional consideration is given when a Marine is performing in a role that is inconsistent with the Marine's grade (USMC, 2000), which mostly applies when the Marine is filling a billet that is typically filled by a Marine of higher grade.

MARK	CORRESPONDING ADJECTIVE RATING	STANDARDS OF PROFICIENCY
0.0 to 1.9	Unacceptable	Does unacceptable work in most duties, generally undependable; needs considerable assistance and close supervision on even the simplest assignment.
2.0 to 2.9	Unsatisfactory	Does acceptable work in some of the duties but cannot be depended upon. Needs assistance and close supervision on all but the simplest assignments.
3.0 to 3.9	Below Average	Handles routine matters acceptably but needs close supervision when performing duties not of a routine nature.
4.0 to 4.4	Average	Can be depended upon to discharge regular duties thoroughly and competently but usually needs assistance in dealing with problems not of a routine nature.
4.5 to 4.8	Excellent	Does excellent work in all regular duties, but needs assistance in dealing with extremely difficult or unusual assignments.
4.9 to 5.0	Outstanding	Does superior work in all duties. Even extremely difficult or unusual assignments can be given with full confidence that they will be handled in a thoroughly competent manner.

Figure 2. Guidance and Standards in Assigning Duty Proficiency Marks. Source: USMC (2000).

2. Conduct Marks

Conduct marks measure how well the Marine conforms to standards and regulations. Marks range from 0.0 to 5.0. Figure 3 describes the conduct marks' corresponding adjective ratings and narratives. According to the IRAM (USMC, 2000), the following is what should be considered when assigning conduct marks:

> General bearing, attitude, interest, reliability, courtesy, cooperation, obedience, adaptability, influence on others, moral fitness, physical fitness as effected by clean and temperate habits, and participation in unit activities not related directly to unit mission. (p. 4–39)

Additional consideration is given if the Marine was assigned to the weight control program during the evaluation period (USMC, 2000).

MARK	CORRESPONDING ADJECTIVE RATING	STANDARDS OF CONDUCT
0.0 to 1.9	Unacceptable	Habitual offender. Conviction by general, special, or more than one summary court-martial. Give a mark of "0" upon declaration of desertion. Ordered to confinement pursuant to sentence of court-martial. Two or more punitive reductions in grade.
2.0 to 2.9	Unsatisfactory	No special court-martial. Not more than one summary court-martial. Not more than two nonjudicial punishments. Punitive reduction in grade.
3.0 to 3.9	Below Average	No court-martial. Not more than one nonjudicial punishment. No favorable impression of the qualities listed in paragraph 4007.6a. Failure to make satisfactory progress while assigned to the weight control or military appearance program. Conduct such as not to impair appreciably one's usefulness or the efficiency of the command, but conduct not sufficient to merit an honorable discharge.
4.O to 4.4	Average	No offenses. No unfavorable impressions as to attitude, interests, cooperation, obedience, after-effects of intemperance, courtesy and consideration, and observance of regulations.
4.5 to 4.8	Excellent	No offense. Positive favorable impressions of the qualities listed in paragraph 4007.6a. Demonstrates reliability, good influence, sobriety, obedience, and industry.
4.9 to 5.0	Outstanding	No offenses. Exhibits to an outstanding degree the qualities listed in paragraph 4007.6a. Observes spirit as well as letter of orders and regulations. Demonstrates positive effect on others by example and persuasion.

Figure 3. Guidance and Standards in Assigning Conduct Marks.
Source: USMC (2000).

B. QUANTITATIVE PERFORMANCE MEASURES

This section lists the quantitative performance measures that are used in combination with proficiency and conduct marks to calculate promotion scores.

1. Rifle Marksmanship Score

The rifle marksmanship score measures the Marine's ability to apply the fundamentals of marksmanship and effectively employ the service rifle in varying conditions (USMC, 2014). Its cultural importance is represented by one of the Marine Corps' enduring principles: "Every Marine is a rifleman" (Dunford, 2015). Additionally, Chung et al. (2011) identify significant relationships between rifle marksmanship and aptitude, psychomotor skills, and affective variables such as anxiety, which suggests that rifle marksmanship represents, to some extent, the Marine's cognitive, non-cognitive and physical abilities.

2. Physical Fitness Test and Combat Fitness Test

The Physical Fitness Test (PFT) and the Combat Fitness Test (CFT) measure the Marine's self-discipline, commitment, and individual combat readiness (USMC, 2008). The PFT and CFT are biannual events—the PFT in the first half of the year and the CFT in the latter half. Recent analysis of the PFT and CFT led the Marine Corps to adjust scoring standards to allow for better differentiation between individual fitness levels (USMC, 2016a).

3. Self-Education

Self-education is intended to measure commitment to intellectual growth (MCU, 2012). Self-education includes self-paced Marine Corps–sponsored education through MarineNet and Marine Corps Institute (MCI) courses, college credits received from traditional and vocational educational institutes, and through the College Level Examination Program (CLEP).

4. Special Duty Assignments

Drill instructor (DI), recruiter, Marine security guard (MSG), combat instructor, and Marine Corps Security Forces (MCSF) are special duty assignments that are filled by Marines from any occupational field (OccFld). A special duty assignment involves demanding duties and an unusual degree of responsibility (USMC, 2001). Marines serving or who have successfully served in a special duty assignment, especially as a DI

or recruiter, are assumed to be highly qualified for promotion (USMC, 2001). The incentives associated with special duty assignments, including extra pay and duty station preference, suggest two things. One is that filling special duty assignment billets with volunteers is difficult. Two, the Marine Corps values the personal traits required to satisfactorily complete an unusually difficult and demanding assignment, which may include resiliency and commitment, among others.

5. Experience/Seniority

Experience is measured by time in grade (TIG) and time in service (TIS). Experience can be a proxy for accumulated human capital that is not tangible and is difficult to measure through changes in performance over a short period of time. Accounting for experience reduces distortion in the incentive scheme, namely promotions. For instance, a Marine may be highly qualified for promotion, but becomes eligible to promote at a time when no vacancies exist at the next higher grade. That Marine has no choice but to wait for vacancies. When vacancies become available, the Marine who was forced to wait receives priority over other similarly qualified yet less experienced Marines. Alternatively, the Marine with relatively low productivity eventually accumulates enough experience points to compete with the highly productive, less experienced Marine.

C. ADMINISTRATIVE FUNCTIONS

Proficiency and conduct marks are used as a basis for several administration functions including promotion, retention, selection to special duty assignments, and characterization of service upon discharge.

1. Promotion—Composite Score

Marines compete for promotion within each grade and primary military occupational specialty (PMOS) based on past performance. The intent of the promotion system is to promote only those who are qualified to assume the responsibilities of the next higher grade (USMC, 2012). Thus, past performance is used as a metric to predict the Marine's potential at the next higher grade. The most basic requirements for

promotion are TIS and TIG. Marines at the rank of private and private first class are promoted to the next higher rank once they meet the required TIG and TIS (requirements may be waived for meritorious promotion [USMC, 2012]). Lance corporals and corporals are promoted based on a composite score. Lance corporals and corporals are ranked within grade and PMOS by something called a composite score. The composite score is an overall "quality" score that combines multiple performance-related elements, seniority, education, and special duty assignments, if applicable.

In addition to obtaining a composite score, Marines must also be eligible for promotion. Eligibility is determined by meeting minimum TIG and TIS and not having received a non-recommendation for promotion (reported as "NOT REC PROM" in the unit diary). It is possible for a Marine to have a composite score and not be eligible for promotion.

The composite score consists of 10 performance-related elements. The first part of the composite score consists of three performance scores grouped into a general military proficiency (GMP) score. The GMP is composed of the rifle marksmanship score, PFT score, and CFT score. Each score is converted to a rating, then averaged to create a total GMP score (Table 3, lines 1 through 6). The score to rating conversion is covered in more detail in Chapter IV. Job-related performance elements are captured by average duty proficiency and conduct marks since last promotion. Seniority is calculated with TIG and TIS. Lastly, additional "bonus" points can be accumulated if the Marine is currently performing or has satisfactorily completed assignment in a special duty assignment since the last promotion. The Marine can also receive bonus points for self-education and for referring individuals to Marine Corps recruiters if the referred individual enlists. Table 3 shows the calculation method and weights of each element.

Table 3. Composite Score Calculation Method and Weights.
Adapted from USMC (2012).

Line No.			Max Possible Score or Rating	Weight of Max Possible
1.	Rifle Marksmanship Score _____	=	5.0	7.2%
2.	PFT Score _____	=	5.0	7.2%
3.	CFT Score _____	=	5.0	7.2%
4.	Subtotal (line 1+2+3)	=	15.0	
5.	GMP Score (line 4 divided by 3)	=	5.0	
6.	GMP Score (from line 5) _____ x 100	=	500.0	
7.	Average Duty Proficiency _____ x 100	=	500.0	21.6%
8.	Average Conduct _____ x 100	=	500.0	21.6%
9.	Time in Grade[a] (months) _____ x 5	=	320.0	13.8%
10.	Time in Service[a] (months) _____ x 2	=	192.0	8.3%
11.	DI/Recruiter/MSG/Combat Instructor/MCSF	=	100.0	4.3%
12.	Self-Education Bonus			
	a. MarineNet/MCI/Extension School _____ x 15	=	60.0	2.6%
	b. College/CLEP/Vocational _____ x 10	=	40.0	1.7%
13.	Command Recruiting Bonus	=	100.0	4.3%
14	Composite Score (sum of lines 6 through 13)	=	2312	100.0%

[a] TIG is based on the Target Enlisted Career Progression Pattern and service limitations for a corporal competing for promotion to sergeant.

Composite scores are automatically computed quarterly once the Marine has met the required TIG and TIS for the promotion quarter. For example,

> a Marine with a [Lance Corporal] date of rank of 1 June 2003 will have served 8 months TIG on 1 February 2004 and will have a composite score computed for the January, February, and March 2004 promotion quarter. If the Marine meets the required cutting score for 1 January 2004, the unit will receive a "SELECT GRADE" on the [Diary Feedback Report]. (USMC, 2012, p. 2–11)

The cutting score is the lowest composite score within grade and PMOS that is authorized for promotion. It is the mechanism that controls the number of Marines to be promoted each month based on vacancies (USMC, 2012). When no vacancies exist

within a PMOS at the promotion grade, the cutting score is reflected as "closed" and no promotions occur.

2. Retention—First-Term Alignment Plan and Computed Tier Score

The relationship between retention and proficiency and conduct marks is not the focus of this study, but it is relevant to better understand the impact of proficiency and conduct marks on a Marine's career.

Much like promotion, reenlistment is a competitive process in which its success is measured by being able to identify only the most deserving Marines. One of the primary goals of the Marine Corps' retention policy is to retain the most qualified Marines by grade and PMOS (USMC, 2010, p. 1–1). The First-Term Alignment Plan (FTAP) is the retention policy used to retain first-term Marines for the purpose of meeting career force requirements by PMOS and preventing promotion stagnation (USMC, 2010). Marines typically compete for a limited number of reenlistment spaces, except for a few PMOSs that have more spaces than reenlistment submissions. Historically, the FTAP reenlistment rate across the Marine Corps is about 24 percent (Crider, 2015). The fiscal year (FY) 2017 FTAP goal is to retain 23 percent of first-term Marines with an end of active service date in FY2017 (USMC, 2016b). Thus, it is critical to give the most deserving Marines first opportunity to fill limited reenlistment spaces.

The computed tier score, introduced in May 2011, is used to identify the most deserving Marines for reenlistment opportunity. The computed tier score is similar to the composite score. Rifle marksmanship score, PFT score, and CFT score are included, though not converted to a ranking as in the composite score. Proficiency and conduct marks are also included. The computed tier score introduces the Marine Corps Martial Arts Program belt level and any meritorious promotions during the enlistment period as additional measures of quality. Table 4 presents the computed tier score method and component weights. The total score is used as the basis to determine placement into one of the four tiers. Tier assignment is relative to all other Marines with the same PMOS and an end of active service date in the same fiscal year (Crider, 2015). Table 5 shows the tier

distribution. Marines charged with misconduct during their enlistment are automatically excluded from Tier I.

Table 4. Computed Tier Score Method and Weights. Adapted from Cole (2014).

Component		Max Possible Score	Weight of Max Possible
Rifle Marksmanship Score	=	350	16.3%
CFT Score	=	300	14.0%
PFT Score	=	300	14.0%
Average Duty Proficiency _____ x 100	=	500	23.3%
Average Conduct _____ x 100	=	500	23.3%
MCMAP Belt Points	=	100	4.7%
Meritorious Promotion	=	100	4.7%
Total	=	2150	100%

Table 5. Tier Distribution within MOS and EAS FY. Adapted from Cole (2014).

Tier	Description	Distribution
1	Eminently Qualified	91% - 100%
2	Highly Competitive	61% - 90%
3	Competitive	11% - 60%
4	Below Average	1% - 10%

3. Competitive Programs

Several competitive programs, such as special duty assignments and enlisted to officer commissioning programs, use performance measures to select the most qualified from an applicant pool. Average duty proficiency and conduct marks are included in the screening process for recruiting, Marine security guard, and independent duty, as well as the commanding officer's endorsement of a Marine applying for an enlisted to officer commissioning program (USMC, 2001, 2015a). The minimum average proficiency and conduct marks are 4.6/4.6 for recruiting duty, 4.2/4.2 for Marine security guard duty, and 4.4/4.4 for independent duty (USMC, 2001). Additional requirements in the screening

process relate specifically to the nature of the requested duty, not all of which are performance related. Waivers for not meeting one or more of the requirements are considered on a case-by-case basis (USMC, 2001).

4. Characterization of Service upon Discharge

Characterizations of separation are honorable, general (under honorable conditions), under other than honorable conditions, bad conduct, dishonorable, and uncharacterized (USMC, 2013). Honorable characterization requires the Marine's average proficiency marks to be 3.0 or higher and average conduct marks to be 4.0 or higher, though exceptions can be made (USMC, 2013).

D. RATER TRAINING

Enlisted leaders receive formal rater training on proficiency and conduct marks from the Marine Corps University's professional military education curricula (published at https://vcepub.tecom.usmc.mil/sites/edcom/epme/default.aspx). Formal training starts at the Corporal's Course and progressively evolves at the Sergeant's Course, and then Career Course. Corporals attending the Corporal's Course are introduced to proficiency and conduct marks' process and procedures. Additionally, corporals are expected to realize the significant impact proficiency and conduct marks have on the Marine's career. Sergeants revisit the same material and are given additional instruction on how to support substandard marks with appropriate documentation. Staff sergeants at the Career Course briefly review proficiency and conduct marks' process and procedure. The focus shifts to administrative oversight and ensuring adherence to reporting occasions. According to the Advanced Course material published on MCU's SharePoint site, gunnery sergeants do not appear to receive formal training on proficiency and conduct marks. This is not alarming, however, because senior enlisted Marines are, by the nature of their rank and experience, the subject matter experts in areas such as assigning proficiency and conduct marks. For the fitness report system, senior enlisted advisors "have the responsibility to assist reporting officials and commanders in completing and processing enlisted fitness reports" (USMC, 2015b, p. 2–4). Senior enlisted advisors have a similar responsibility for

the assignment of proficiency and conduct marks (R. VanOostrom, personal communication, February 17, 2017).

E. SUMMARY

Proficiency and conduct marks have the potential to alter a Marine's career. In combination with several quantitative performance measures, proficiency and conduct marks play a major role in several administrative decisions, including promotion and retention. Additionally, proficiency and conduct marks are part of the selection criteria for special duty programs. Raters begin to receive training at the rank of corporal. Raters learn how to assign proficiency and conduct marks and they learn that proficiency and conduct marks, if improperly assigned, can severely alter a Marine's career progression.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LITERATURE REVIEW

In this chapter, I review the academic literature on performance evaluation systems in various types of organizations over the past century. The majority of studies from the early 1900s to circa 1980 focused on improving the psychometric quality of the rating procedures pertaining to usefulness for pay and promotion decisions (e.g., Jacobs, Kafry, & Zedeck, 1980; Landy & Farr, 1980; Lawshe, 1975; Saal, Downey, & Lahey, 1980). Much of the recent literature (since circa 1980) focuses on how to use performance appraisals as a feedback and development tool to boost employee productivity and increase job satisfaction (e.g., Cederblom & Pemerl, 2002; Locke & Latham, 1990; Longenecker, Liverpool, & Wilson, 1988). This study focuses on the former literature that pertains to the effectiveness of the performance appraisal procedures used as a basis for administrative decisions.

A. LABOR ECONOMIC THEORY: INTERNAL LABOR MARKETS

Internal labor markets enable human capital investment, induce worker productivity, and also depend on an effective performance evaluation system to thrive. In this section, I describe why the military manpower system fits the mold of an internal labor market. Then, I introduce human capital theory and follow with a discussion of the promotion tournament model. This section provides a broad understanding of labor activity within the military and supports this study's relevancy.

Internal labor markets have two distinct properties: (1) there are limited points of entry and exit, and (2) labor allocation and wage determination are governed by administrative rules and not by external economic fluctuations (Doeringer & Piore, 1985). Limited entry and exit points mean that movement into the labor market occurs at certain job classifications, and promoting or transferring workers who have already gained entry fill all other positions (Doeringer & Piore, 1985). The single entry point for Marine enlistees is through recruit training. Following graduation of recruit training and earning the title "Marine," enlistees are classified into a PMOS. Marines are then promoted to fill more senior positions within their PMOS. Marines who have already

gained entry to the labor market may occasionally transfer to fill vacancies in a different PMOS.

Labor allocation and wage determination are not directly influenced by external market forces (Doeringer & Piore, 1985). Fluctuations in the external market labor supply will not directly influence the promotion rules of the internal labor market because jobs are filled internally. The Marine Corps' promotion system is relatively stable, although fluctuations in external market conditions may impact retention rates and call for an increased number of promotions to fill vacated positions.

1. Firm-Specific Human Capital

There are two types of human capital a worker can accumulate: general training and specific training. General training is equally valuable inside and outside the firm (Lazear & Gibbs, 2015). Specific training, also called firm-specific human capital, is training that increases productivity at the current firm but does not raise the worker's value to other firms (Lazear & Gibbs, 2015). In reality, most training falls somewhere in between the two, and most firm-specific training will still provide some value to other firms (Lazear & Gibbs, 2015).

The internal labor market's rigid structure enables firm-specific human capital investment (Doeringer & Piore, 1985). In competitive markets, workers do not invest in specific training for fear of involuntary unemployment, which would result in the worker suffering the cost of the training investment. Also, employers are not incentivized to pay for specific training for fear that the worker will leave before the investment is recouped. In internal labor markets, the worker is encouraged to invest in firm-specific human capital because of employment guarantees tied to promotion and retention rules (Doeringer & Piore, 1985).

2. Promotions: The Tournament Model

Internal labor markets give rise to tournament promotion systems, which allow firms to fill vacancies with top performers. Tournaments are similar to a sports tournament or playoff system in which teams or players advance if their performance is

better than their opponents'. The margin of victory is not important, only that the winner's score is higher than the loser's. In labor market terms, tournaments exists when a fixed number of workers is promoted based on their relative, instead of absolute, performance (Lazear & Gibbs, 2015). If tournament rules did not exist, then any number of workers would be promoted upon reaching a pre-defined performance level (Lazear & Gibbs, 2015). It makes sense for tournaments to appear in the internal labor market where a number of workers who have already gained entry to the market are competing for a limited number of promotion spots.

Tournaments have several advantages for a firm. Tournaments make evaluations and promotion decisions easier to determine because the firm needs only to identify relative performance (Lazear & Gibbs, 2015). This reduces the effect of measurement error in the evaluation process. Additionally, because it is usually apparent who the top performers are, workers may feel that promotion decisions are objective (Lazear & Gibbs, 2015). Another advantage is that the number of promotions is controlled, which prevents overcrowding at certain grades or, alternatively, excess vacancies.

In spite of the advantages, tournaments have their disadvantages as well. Tournaments discourage cooperation among workers, ignore absolute quality, and may not account for factors that affect a worker's performance and are beyond the worker's control, such as local market conditions or different management styles across the organization. First, relative evaluations do not incentivize workers to cooperate because the worker is interested in out-performing his or her co-workers rather than helping boost the co-workers' performance (Lazear & Gibbs, 2015). Fortunately, the incentive structure may be changed to address any potential sabotage. Second, tournaments do not effectively control for quality (Lazear & Gibbs, 2015). The top performers among a pool of lower quality workers are still promoted (Lazear & Gibbs, 2015). Alternatively, high quality workers may not be promoted if the promotion pool contains an unusually high number of quality workers (Lazear & Gibbs, 2015). Quality may also vary when the numbers of promotions vary. For example, USMC officer quality during force buildup, when promotion and retention rates were high, was lower than officer quality during force drawdown, when promotion and retention rates were low (Griner, 2016). Lastly,

tournaments, and relative performance evaluations, may not sufficiently account for factors that affect performance and are beyond the worker's control.

In summary, the tournament promotion system exists in internal labor markets because the most qualified within a pool of eligible workers are promoted to fill the limited number of vacancies. That also means promotions are very much a sorting mechanism that depends on an effective performance evaluation system to identify top performers with the greatest potential to succeed at the next higher grade.

B. FACTORS INFLUENCING PERFORMANCE APPRAISAL EFFECTIVENESS

The worker's performance should be measured in a way that best translates their behaviors and actions into a quantifiable assessment of the worker's performance (Landy & Farr, 1983). Typically, some form of performance appraisal is helpful in making that translation. Of course, as with any other method, performance appraisals are imperfect. The purpose of this section is to discuss the factors that affect the quality and usefulness of the performance appraisal and ultimately to build support for methodology and analysis later in this study.

Figure 4 represents the factors that influence performance measurement. Work performance itself is affected by the worker's environment and individual characteristics (Landy & Farr, 1983; Hosek & Mattock, 2003). The internal and external environment may include peers, command climate, and operational tempo, as well as global factors that affect all organizations. Individual characteristics include ability and motivation. Next, an individual's work performance must be observed, evaluated, and recorded. The performance measurement procedures include performance measures, rating format, and the rater, which determine how accurately the worker's performance is translated into a performance score (Landy & Farr, 1983; MacDonald & Sulsky, 2009). The focus here is on the performance measures, rating format, and rater.

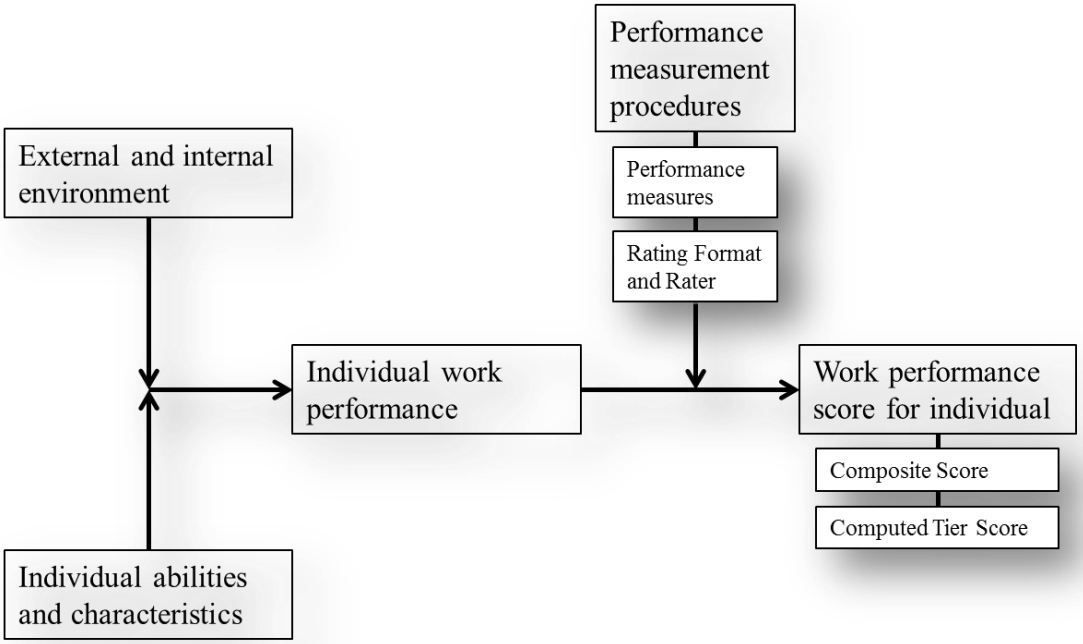

Figure 4. Factors Influencing the Measurement of Work Performance. Adapted from Landy & Farr (1983).

1. Performance Measures

An effective performance measure is one that can be observed and interpreted with which the rater can formulate a reliable judgement of work performance (Smith, 1976). Individual work performance can be classified as the worker's actual behavior, the work outcomes (results) of their behavior, or their overall contribution to organizational goals (Smith, 1976). The latter is the ideal criterion, yet it is practically difficult to measure (Smith, 1976). Thus, performance criteria are usually the behaviors and results that are deemed important to organizational effectiveness (Landy & Farr, 1983; Smith, 1976).

Effective performance measures, whether of behavior or results, have basic requirements: reliability, validity, accuracy (Landy & Farr, 1983), and practicality (Jacobs et al., 1980). Each requirement applies to objective and subjective performance measures, although accuracy and practicality tend to relate more to subjective measures.

Moreover, there are advantages to using subjective performance measures. Subjectivity allows evaluators to remove uncontrollable factors affecting performance, which can make the system more flexible, less distorted, and thus more effective (Lazear & Gibbs, 2015). On the other hand, administrators, raters, and ratees might be dissatisfied with subjective measures because they are prone to intentional and unintentional bias (Landy & Farr, 1980). Administrators should seek to leverage the advantages of subjective measures while minimizing the bias, particularly as it pertains to the research questions explored in this study.

a. Reliability

Reliability means that a performance measure is consistent and stable (Hutchinson, 2013; Landy & Farr, 1983; Smith, 1976). Landy and Farr (1983) define reliability as "the extent to which a set of measurements is free from variance due to random error or the extent to which the variance in a set of measurements is due to systematic sources" (p. 9). It is an important contributor to the quality of the rating procedure (Jacobs et al., 1980). Reliability is a term that is considered in many applications ranging from academic assessments to psychological tests. This study borrows the quantitative tests for stability and interrater reliability, because they are the types of reliability most relevant to performance evaluations.

Stability refers to the amount that performance measurements remain stable over time. Stability is measured by the test–retest method that compares ratings on a group of individuals at one time with ratings on the same group at a later time. Several factors that influence the accuracy of stability estimates need to be considered. The time period between evaluations is one of the more significant influences. Increased duration of observation improves reliability because short-term randomness in performance is "smoothed out" over time. Alternatively, shorter durations may reveal high reliability estimates because of the rater's tendency to anchor the present evaluation to previous marks (Landy & Farr, 1983). Stability estimates may also be affected by the crudity of the measurements. Measures that are not sensitive to actual performance change may appear to be stable (Landy & Farr, 1983).

Interrater reliability is the amount of rater variation on one individual or one level of performance. High consistency means that the same worker receives comparable evaluations at a single point in time from more than one evaluator with whom the worker has similar relationships (Jacobs et al., 1980). Due to data limitations of not having individual rater information, this study estimates reliability at an aggregate level (squadron or battalion level) across different points in time.

b. Validity

Validity means that the performance measure is measuring what it is intended to measure (Hutchinson, 2013; Landy & Farr, 1983) and also that what is intended to be measured is relevant to the organization's goals. As opposed to reliability, which provides information on the behaviors and results that can cause variance in performance measures, validity provides information on the relevancy of those behaviors and results for the purpose of measurement. The most pertinent types of validity for this study are predictive validity and construct validity.

The primary purpose of the Marine Corps' performance evaluation system is to identify Marines who are most qualified to perform at the next higher grade. Therefore, I am concerned with the criterion's predictive validity and how well current measures predict performance at a later time. This study tests for predictive validity by comparing the relationship between proficiency and conduct marks and fitness report scores at a later time, when both measures can be observed in a Marine's career, as explained further below.

Construct validity is concerned with how well the criterion measures the behavior it is intended to measure. This applies when a criterion is supposed to measure a non-observable attribute or characteristic by inferring its value based on an observable performance behavior (Landy & Farr, 1983). For instance, the rater may indirectly observe that proficiency is indirectly observed through one's task completion and perceived effectiveness in achieving results. Conduct is inferred through occasions of misconduct and how well the Marine exhibits favorable characteristics, as defined in the IRAM.

c. Accuracy

Accuracy is concerned with matching as closely as possible the rating and the true level of performance. A rating that is consistently inflated may have high reliability and validity and not be accurate. Accuracy errors prevent performance evaluations from discriminating between individuals and thus make it difficult for decision-makers to separate *true* high performers from *true* low performers. One way to estimate a rating procedure's accuracy is to estimate its susceptibility to rater errors and biases (Jacobs et al., 1980). These errors include halo, leniency, and central tendency.

Halo error is the rater's tendency to anchor ratings of one characteristic closely to the rating of a different characteristic (Smith, 1976), or it is when ratings of all characteristics are based on some global impression of each ratee (Saal et al., 1980). Halo error prevents ratings from reflecting the true level of performance in each measure.

Logical error is another that leads to similar ratings across dimensions (Jacobs et al., 1980; Smith, 1976). Logical errors arise when the rater cannot distinguish the difference between two or more performance dimensions. For example, *leadership* and *influence on others* are expressions used to describe proficiency and conduct marks, respectively. The rater may interpret the two expressions as having the same meaning and assign similar marks for proficiency and conduct. I cover logical error in more detail at the end of this chapter where I form hypotheses about the accuracy of proficiency and conduct marks.

Leniency error is the tendency for raters to give ratings higher than what is deserved (Jacobs et al., 1980). The *true* performance mean is displaced toward the high end of the scale. The opposite displacement may occur as result of severity. There are many reasons for raters to be lenient or severe. One is general human kindness and the hesitancy to rate individuals below average. Another is that strict raters who scrutinize performance more closely may have a higher standard of performance.

Central tendency is the avoidance of ratings at the extreme ends of the scale (Landy & Farr, 1983). The majority of the ratings are grouped at the center of the rating scale. A similar error, called restriction of range, is when the same kind of grouping

occurs somewhere other than the middle of the scale. In either case, the result is a failure to discriminate (Smith, 1976).

Accuracy is affected by the frequency of ratings, much like estimates of stability, and the cognitive aspects of performance ratings. Long periods between ratings will introduce rater error due to selective recall (Smith, 1976). Short periods will not allow raters to observe significant changes in performance. A possible compromise is to encourage performance diaries (also known as training jackets) and to improve rater training (Landy & Farr, 1983). However, as discussed in the next section, the compromise needs to be practical. It is not practical to require continuous recording of worker behavior, and it may not be practical to require comprehensive training for raters. Rater training, nonetheless, seems to be the fix for accuracy problems, although an improved rating format has its role as well.

Achieving accuracy in the rating process is primarily limited by the cognitive component. Cognitive theory is beyond the scope of this research, yet it is important to recognize that, all else equal, the human cognitive limitations prevent a perfect translation of true performance into a performance score.

d. Practicality

Practicality means that performance measures are observable, interpretable, usable, and acceptable to those who need it to make personnel decisions (Smith, 1976). Practicality considers the cost of development, implementation, and execution compared to the potential benefits of improving reliability, validity, and accuracy. Practicality also means that the rating procedures are not overly burdensome to stakeholders. For example, it is conceptually impractical to use fitness report procedures for the junior enlisted force. Completing fitness reports is a tedious administrative task that would be too burdensome to apply to such a large population (E4s and below make up about 60 percent of the total force [computed from active component tables; Center for Naval Analyses, 2015, pp. 92, 96]).

2. Rating Format

The type of rating format depends on the type and purpose of performance measure. Performance measures can be either subjective or objective, and absolute or relative. Absolute methods are based on a standard of performance without direct reference to others, are costly to develop and administer, and are useful in larger organizations (Hutchinson, 2013). Relative measures seek to rank employees based on relative performance, which is an effective way to differentiate between employees for administrative decisions such as promotion and retention and work best in smaller organizations (Hutchinson, 2013). Absolute methods take the form of ratings, and relative methods take the form of comparisons or rankings. This section briefly discusses the various types of ratings to provide a reference for further evaluation later in this research. The principal types of ratings discussed here are graphic rating scales, behaviorally anchored rating scales, and forced choice rating scales.

a. Graphic Rating Scales

Graphic rating scales are a trait-based evaluation tool. They generally consist of a briefly defined trait label for each dimension, which is anchored by an adjective corresponding to each level of performance along the scale (Landy & Farr, 1983). The ends of the scale delineate the extreme levels of performance (e.g., unacceptable to outstanding). Graphic rating scales come in many shapes and sizes. Scales can range from three to seven or more levels, the five-point scale being the most common, though there is no evidence that supports one scale size over another (Hutchinson, 2013). Anchors may be numerical or adjectival, and adjectives may or may not be supported by a definition (Landy & Farr, 1983). Graphic rating scales should be as specific as needed to allow the rater to easily interpret and understand its content in order to improve interrater reliability.

The primary advantage and disadvantage of graphic rating scales is their non-job-specific nature. They can be widely applied across myriad job types and are inexpensive to develop and administer (Wiese & Buckley, 1998). Alternatively, trait-based performance dimensions are difficult to measure and require the rater to infer personality traits from

observed behavior (Wiese & Buckley, 1998). In addition, graphic rating scales are prone to central tendency, leniency, and inconsistency between raters (Hutchinson, 2013).

b. Behaviorally Anchored Rating Scales

Behaviorally anchored rating scales (BARS) were devised to improve the appraisal's psychometric quality by focusing on specific job behaviors rather than ambiguous traits. BARS are constructed similarly to graphic rating scales except that the numeric or adjectival anchors are replaced with behavioral definitions of job-specific work behaviors. Conceptually, a specific definition of work performance makes the appraisal easier to interpret and therefore improves interrater reliability and reduces bias (Hutchinson, 2013). BARS development is a lengthy and arduous process that involves many job experts along the multiple stages of development to form behavior examples, and to define and test the behavioral dimensions for scale placement (Landy & Farr, 1983). The thorough design process is intended to improve the appraisal's validity (Hutchinson, 2013).

An advantage of BARS, besides improving psychometric quality, is that they are amenable to expectation format. Expectation format means that behavioral illustrations can be expressed in terms of predicted behavior (Landy & Farr, 1983). That is, the rater uses observed past performance of the worker to make predictions on the worker's expected performance level in the future as defined by the anchors (Landy & Farr, 1983). This is particularly useful for appraisals that are used as a basis for promotion decisions.

The primary disadvantages of BARS are its costly development and the likely inability of the rater to observe each of the narrowly defined job behaviors (Landy & Farr, 1983). Despite significant efforts to improve the psychometric quality of the rating process, BARS are not proven to be substantially better than graphic rating scales (Hutchinson, 2013; Landy & Farr, 1983; Wiese & Buckley, 1998).

c. Forced Choice Rating Scales

Forced choice is another method aimed at reducing rater errors. It was initially developed by the U.S. Army in the 1940s to correct for leniency and central tendency

bias produced by their graphic rating scale (Landy & Farr, 1983). Forced choice asks the rater to select from a list of job performance examples that best describe the worker (Landy & Farr, 1983). The value of each performance example is determined by job experts and through prior research based on how favorable the item and by how much it discriminates between high and low performers (Landy & Farr, 1983). The item's favorability and discriminatory index are not disclosed to the rater, thus preventing the rater from assigning inflated scores (Landy & Farr, 1983). Forced choice simplifies the rating process because it does not require raters to match observed behaviors with traits, as with graphic rating scales (Wiese & Buckley, 1998). However, raters are resistant to this method because they do not always agree with the given choices nor do they like the secrecy and inability to control the overall rating (Wiese & Buckley, 1998).

d. Summary

Absolute performance measures, such as ratings, are necessary for large organizations when relative ranking methods are simply infeasible. Graphic rating scales are trait-based performance measures that lack interrater reliability, validity, and accuracy. BARS and forced choice are behavioral-based performance measures that improve interpretability and translation of observed behavior to a performance score. BARS and forced choice have associated costs, however, either with development or with rater resistance. In addition, the sophisticated behavioral-based methods have not proven to consistently outperform a well-designed graphic rating scale. It is uncertain which performance appraisal method is best. Jacobs et al. (1980) suggest that rater training may be the better way to reduce errors in the appraisal system (p. 630).

3. Rater

As previously mentioned, the rater is another factor influencing the effectiveness of the performance appraisal. The rater can introduce errors into the performance appraisal (e.g., halo, leniency, central tendency, selective recall). Rater-induced errors are influenced by the rater's ability to observe those who are being evaluated as well as intentional and unintentional biases.

a. Who Should Evaluate?

The primary requirement of a rater is that they have the opportunity to observe the behavior of the individual being evaluated (Smith, 1976). Additionally, the rater should have the expertise and job-specific knowledge upon which to base their judgment of work performance (Lazear & Gibbs, 2015). For these reasons, the rater is quite often the immediate supervisor. The rater could also be a more remote superior, peer, subordinate, or self. A worker may also have multiple evaluators. Multiple evaluators will reduce the likelihood that the final evaluation is biased (Lazear & Gibbs, 2015), most likely, because different biases of different raters effectually cancel each other out.

b. Training Related to Unintentional Bias

Scholars tend to agree that rater training is more significant than rating format in determining the success of the performance appraisal procedures (e.g., Hutchinson, 2013; MacDonald & Sulsky, 2009; Sulsky & Day, 1992; Smith, 1976). Smith (1976) also identifies that rater ability, intelligence, and relationship with subordinate are factors influencing rater quality. Nonetheless, the focus on improving rater quality seems to revolve around improving cognitive processes. Surprisingly or not, rater accuracy is improved by simply teaching raters how to avoid cognitive-related biases (Pursell et al., 1980).

Frame of reference (FOR) training is another type of training that is shown to significantly improve rater accuracy. It is based on the assumptions that raters form general impressions of their subordinates over time rather than use specific behavioral information (Sulsky & Day, 1992). FOR training is designed to give all raters a common reference for different levels of performance (Sulsky & Day, 1992). Raters build a common understanding of what job behaviors belong at each performance level, which ultimately results in the rater forming appropriately classified general impressions of their subordinates.

c. Training Related to Intentional Bias

Up to this point, I have discussed unintentional biases that result from poorly constructed rating formats or rater ignorance. Intentional bias is also worthy of mention because of its potential to deteriorate rating effectiveness. Intentional bias is determined by the extent that the rater is motivated to make accurate assessments of their subordinates (MacDonald & Sulsky, 2009). Biases can be positive, negative, or indifferent and may be influenced by favoritism, race, gender, and the rater's attitude toward the rating process. For instance, Landy & Farr (1983) reveal that males receive more favorable evaluations than females in jobs that are traditionally performed by males, all else being equal. MacDonald & Sulsky (2009) emphasize the importance of rater training and stress that it should focus foremost on the factors affecting rater behavior, followed by improving rating accuracy.

4. Summary

There is no ideal rating format. Actual performance is effectively translated into a performance score when performance appraisal procedures maximize reliability, validity, accuracy, and practicality. Landy & Farr (1983) assert that the solution is a compromise between methods:

> Dimension labels may be trait names whereas the dimension definitions and scale anchors may be task and behavior oriented. This may allow us the objectivity of measurement we desire while providing a form of relatively standardized information about traits. The operational definition of traits in terms of job behaviors does not perfectly solve the content issue but it seems better than a strict reliance on just traits or just job behaviors. (p. 87)

The actual format is less important than ensuring that its complexity is consistent with the rater's abilities. A well-designed rating format is only as good as the rater who is using it. Rating accuracy can be improved with training, either through making raters aware of potential biases or by assisting them with categorization of performance levels.

Lastly, the psychometric properties of the performance evaluation procedures should be evaluated in the context of the purpose it serves. Procedures designed to

provide information for the purpose of promotion and retention will have much different validity than procedures designed to increase worker productivity.

C. PREVIOUS STUDIES

In this section, I highlight two studies conducted specifically on proficiency and conduct marks as well as two other studies that are relevant to this research. Of the first two studies, Headquarters Marine Corps conducted one in 1996, and the Center for Naval Analyses conducted the other in 1986. Of the other two studies, one includes proficiency and conduct marks as valid measures of Marine enlistee quality (Crider, 2015), and the other focuses on the effectiveness of the fitness report, the Marine Corps' other performance appraisal method (Clemens, Malone, Phillips, & Lee, 2012).

1. Headquarters Marine Corps Study in 1996

In 1996, Headquarters Marine Corps (HQMC) conducted a study that addressed perceived issues of proficiency and conduct mark inflation and their lack of administrative usefulness (W. Wathen, personal communication, October 25, 2016). The study analyzed the active duty lance corporal and corporal population at the time (about 61,000 Marines). Results suggested the presence of inflation, although with little harm to the marks' administrative usefulness.

HQMC analyzed the distribution of proficiency and conduct marks in comparison with the intended distribution per the IRAM. A good (average) Marine, according to the IRAM, should receive marks within the 4.0 to 4.4 range (USMC, 2000). The results showed that 55 percent of corporals had average in grade proficiency marks in the 4.6 to 4.7 range. About 55 percent of lance corporals had average in grade proficiency marks in the 4.5 to 4.6 range. Results were similar, though slightly lower, for conduct marks. Additionally, the distribution of marks varied by OccFld. The 03 OccFld (combat arms) received much lower marks on average than the 01 OccFld (administration) and 60/61 OccFlds (aircraft maintenance). The differences between OccFld were deemed acceptable, however, because promotion decisions were made within each MOS. The study ultimately concludes that proficiency and conduct marks are relevant indicators of proficiency and conduct (W. Wathen, personal communication, October 25, 2016).

2. Mayberry (1986)

Mayberry's (1986) study suggests that proficiency marks are a valid measure of performance to the extent that they sufficiently differentiate between levels of performance. Mayberry (1986) compared individual performance differences measured by proficiency marks with the performance differences in two objective measures: hands-on job tests and industrial productivity. Results show that proficiency marks differentiate between performance levels at least the same if not better than the two objective measures.

Mayberry (1986) surveyed 218 Marine Corps officers asking them to place a percentage between proficiency marks indicating the increase in value to the Marine Corps that a Marine with one mark would have over a Marine with the next lowest mark. The survey results were converted into relative-values at each proficiency mark in order to determine percent differences between any two marks. They found that a Marine in the 95th percentile is 161 percent more valuable than a Marine in the 5th percentile. That number resonates a bit more when compared to a 127 percent difference between the 95th and 5th percentile scores for hands-on job tests and a 106 percent for industrial labor productivity. Although Mayberry's (1986) purpose and results differ from our study, it does tell us that the current system (presumed to be unchanged since before 1981) was, at that time, able to sufficiently differentiate between performance levels.

3. Crider (2015)

Crider (2015) finds that proficiency and conduct marks are valid predictors of future success. The study evaluates all the components of the computed tier score used to guide the administration of first-term reenlistments. The data includes 317,468 Marines who joined the Marine Corps between FY1995 and FY2009. The model he used to analyze the effect of proficiency and conduct marks on future success outcomes includes the computed tier components, fiscal year of reenlistment fixed effects, and PMOS fixed effects. Additionally, he standardized the components to have a mean of zero in order to ease interpretation of differently scaled items (Crider, 2015). Figure 5 depicts his results.

Variables	Months to E6	Months to E7	Stay 6	Stay 8	Stay 10	Stay 12	PFT Reenl+ 2yrs	RelVal Cumulative	ROCV Cumulative
PFT Score	-0.473***	-0.221	0.047***	0.049***	0.058***	0.054***	15.667***	0.341***	0.096***
	(0.089)	(0.197)	(0.002)	(0.003)	(0.003)	(0.004)	(0.234)	(0.022)	(0.005)
Rifle	-0.304**	-0.247	0.016***	0.019***	0.025***	0.021**	0.424**	0.133***	0.033***
	(0.140)	(0.410)	(0.004)	(0.005)	(0.007)	(0.009)	(0.188)	(0.039)	(0.008)
Proficieny	-0.746***	-0.805**	0.012***	0.006	0.011**	0.011	3.421***	0.830***	0.164***
	(0.150)	(0.335)	(0.004)	(0.005)	(0.006)	(0.007)	(0.238)	(0.041)	(0.009)
Conduct	-0.659***	0.053	0.034***	0.038***	0.037***	0.035***	0.391*	0.229***	0.052***
	(0.147)	(0.335)	(0.004)	(0.005)	(0.005)	(0.007)	(0.231)	(0.039)	(0.009)
Merit. Prom.	-0.389***	-0.330**	0.011***	0.010***	0.012***	0.015***	1.666***	0.240***	0.050***
	(0.065)	(0.146)	(0.002)	(0.003)	(0.003)	(0.004)	(0.108)	(0.019)	(0.004)
Observations	25,249	6,168	51,440	35,081	24,703	14,515	62,662	36,214	36,999
R-squared	0.319	0.331	0.075	0.068	0.075	0.075	0.321	0.137	0.131
The model includes a constant and controls for fiscal year of reenlistment and PMOS.									
Robust standard errors in parentheses									
*** p<0.01, **p<0.05, *p<0.1									

Figure 5. The Effects of Standardized Quality Score Components on the Success Outcome. Source: Crider (2015).

The effect of proficiency and conduct marks on fitness report scores (*RelVal Cumulative* and *ROCV Cumulative*) is significant. Crider (2015) gives the interpretation of the standardized score's effect:

> A one standard deviation change or a 0.13 point increase in the proficiency marking is predicted to increase the [reporting senior relative value] cumulative average by 0.83 points. A one standard deviation change or a 0.13 point increase in the proficiency marking is predicted to increase the [reviewing officer cumulative value] cumulative average 0.16 points. (p. 56)

Our research employs a similar model using the first two years of fitness report data on Marines who promoted to sergeant in FY2012 and FY2013.

4. Clemens et al. (2012)

The Director of Manpower Management for the Marine Corps requested for the Center for Naval Analyses to conduct the Clemens et al. (2012) study to check the fitness report system for inflation, fairness, and effectiveness. To check for inflation, the authors use descriptive statistics to show fitness report averages and standard deviations over time. To check for fairness, the authors test for differences in race, gender, and

occupational field as well as differences when the Marine reported on is a different race, gender, or occupational field than the rater. Most relevant to our research are their findings that logistics officers receive higher marks when their rater is also a logistician. Our study similarly tests for interrater reliability. Though we do not have data on rater PMOS, our study uses select occupational fields that are likely to receive proficiency and conduct marks from a supervisor that is of a different PMOS. Additionally, we look for evidence of different grading philosophies across units.

Clemens et al. (2012) find that fitness reports are generally effective in measuring performance. They did recommend, however, that raters receive more substantial training on the performance evaluation system.

D. QUALITATIVE REVIEW OF PROFICIENCY AND CONDUCT MARKS

This section consolidates the broad research presented in the literature review to form a more explicit understanding of the factors affecting proficiency and conduct marks. This section briefly reviews labor economic theory and measures of performance appraisal effectiveness, and then concludes with my hypotheses on the effectiveness of the proficiency and conduct marks.

1. Summary of Literature Review

Internal labor markets create an environment that induces workers to acquire firm-specific human capital in order to increase individual productivity, and to ultimately be rewarded with one of the limited promotion spots. The Marine Corps attempts to differentiate Marines by the amount of accumulated human capital. The relevant measures of human capital in terms of promotion are rifle score, physical fitness scores, self-education, and other quantifiable performance measures. In addition, I assess proficiency and conduct marks to be subjective measures of the Marine's human capital. Furthermore, proficiency and conduct marks attempt to measure the extent to which a Marine pursues Marine-like qualities and to which a Marine's actions positively contribute to the organization.

The information provided by proficiency and conduct marks should be a reliable, valid, accurate, and practical translation of the Marine's true performance. The effectiveness of the translation is affected by the rating format and the rater. The next section provides hypotheses on the effectiveness of proficiency and conduct marks based on the rating format and the rater.

2. Hypothesized Effectiveness of Proficiency and Conduct Marks

I expect proficiency and conduct marks to have low estimates of reliability and accuracy because of the inherent flaws of a graphic rating scale. However, I expect the marks to have high estimates of validity because of the results reported by Crider (2015) that proficiency marks are a significant predictor for future fitness report scores.

Proficiency and conduct marks are subjective, absolute performance measures on a graphic rating scale. As previously discussed, graphic rating scales are prone to central tendency, leniency, and inconsistency between raters (Hutchinson, 2013). That is, I expect proficiency and conduct marks to have low accuracy estimates as well as low interrater reliability estimates. I expect marks to be tightly distributed around the mean and for the actual mean to be higher than the intended mean of 4.2. Proficiency and conduct have the following characteristics in common with a graphic rating scale:

- Performance is measured on a scale from 0.0–5.0.
- The scale is anchored by six adjectives with a corresponding range of values and narratives.
- The marks' definitions are a list of attributes and the anchor definitions are non-job-specific performance examples.

Also related to accuracy, proficiency and conduct marks are likely susceptible to logical error—where the rater is likely to assign similar marks for proficiency and conduct because of the difficulty in interpreting the marks as different. Table 6 lists the attributes listed in the IRAM under each mark (USMC, 2000). Several of the attributes between the two marks may be similarly interpreted by the rater.

Table 6. Attributes Considered in Assignment of Proficiency and Conduct Marks. Adapted from USMC (2000).

	Duty Proficiency Marks	Conduct Marks
Primary attributes	Technical skills Specialized knowledge	Observance of the letter of law and regulations Conformance to accepted usage and custom Positive contributions to unit and Corps
Additional attributes	**Mission accomplishment** **Leadership** Intellect and wisdom **Individual character** **Physical fitness** **Personal appearance** Professional Military Education Marine Corps Institute courses Off-duty education	**Bearing** Attitude **Interest** **Reliability** Courtesy Cooperation Obedience Adaptability **Influence on others** Moral fitness **Physical fitness as affected by clean and temperate habits** Participation in unit activities not directly related to unit mission Assignment to weight control

Note. Bold font indicates the attributes that appear to be related to both proficiency and conduct.

IV. DATA AND METHODOLOGY

A. DATA SOURCES

This research uses data from two sources, Total Force Data Warehouse (TFDW) and Manpower Management Records and Performance Branch (MMRP). TFDW data are semi-annual snapshots of every Marine at the paygrades E1 to E4 from February 2006 to August 2016. The TFDW data are 2,500,656 observations, which represent 418,369 distinct Marines. The average number of observations per Marine is nearly eight. TFDW data include individual performance, demographic, and occupational variables. MMRP data are 26,358 distinct individuals who promoted to sergeant in FY2010 through FY2013 and appear in the TFDW data. MMRP data include fitness report scores for each individual up to three years following the promotion fiscal year.

B. DATA CLEANING AND CODING

Prior to analysis, I clean the data and create new variables in order to simplify nominal data and to correct errors related to unintentional missing values.

1. Proficiency and Conduct Marks

The data include average proficiency and conduct marks in grade that appear at each TFDW snapshot. Eighteen percent, or 460,656, of the observations are missing average proficiency and conduct marks in grade. Of the missing observations, I am able to replace less than 1 percent, or 265, with previous average in grade marks if the Marine was eligible for a composite score and would most likely have submitted a remedial promotion request for a miscalculated composite score per the promotion manual (USMC, 2012, p. 2–28). About 97 percent of the total missing values occur at the Marine's first observation in the data or immediately following promotion.

2. Physical Fitness Test and Combat Fitness Test Scores

I attempt to correct all missing or zeroed PFT scores to the score that would have otherwise been administratively calculated for remedial promotion purposes. Specifically, I correct partial, medical, and combat coded PFTs that show an overall score of zero to

reflect the most recent score attained during the preceding period (per the guidance in the Enlisted Promotion Manual [USMC, 2012]). I am not able to calculate PFT scores prior to 2010 because individual event scores are not available.

Missing CFT data prior to 2010 restricted composite score analysis to years 2010 to 2016. The CFT was first implemented in August 2009 for the period July 2009 through December 2009. However, 97 percent are missing scores for this period, which suggests that the recording procedures did not standardize until the following year. The number of missing scores drops to 23 percent for the same period in 2010, and falls to under 10 percent for the following years. I exclude CFT scores from the reliability analysis portion of this study in order to include years 2006 through 2010 in the analysis.

For analysis of composite score components, I convert PFT and CFT scores to a common rating (USMC, 2012, pp. 2–31–2-32). The conversions are provided in Table 7.

Table 7. PFT and CFT Score to Rating Conversion Table. Adapted from USMC (2012).

Scores		Rating
PFT	CFT	
300	300	5.0
285-299	294-299	4.9
270-284	288-293	4.8
255-269	282-287	4.7
240-254	276-281	4.6
225-239	270-275	4.5
215-224	261-269	4.4
205-214	252-260	4.3
195-204	243-251	4.2
185-194	234-242	4.1
175-184	225-233	4.0
167-174	218-224	3.9
159-166	211-217	3.8
150-158	204-210	3.7
143-149	197-203	3.6
135-142	190-196	3.5
-	-	3.4
110-134[a]	-	3.0
0-134[b]	0-189	0.0

[a] Applies to Marines ages 27 and older
[b] Applies to Marines ages 17–26

3. Rifle Marksmanship Scoring Procedures

Three different scoring systems for rifle marksmanship appear in the data. Prior to 2008, 250-point and 65-point scales were used. In FY2008, the Marine Corps transitioned to a 350-point scale. About 68 percent of the observations have qualifying scores under the 350-point scale. The 350-point scale aggregates two courses of fire: fundamental rifle marksmanship (FRM) and basic combat rifle marksmanship (BCRM), and are more commonly referred to as Table 1 and Table 2. The bimodal distribution of scores in FY2008 indicates that the transition to the new system endured throughout FY2008. 30 September 2008 is the latest date at which a 250-point scale score was recorded. Changes to the rifle marksmanship order (MCO 3574.2K; USMC, 2007) since 2008 have no effect on the scoring procedures.

Less than 1 percent of the observations failed to receive a qualifying score and are coded as "unqualified." About 6.5 percent have missing rifle scores. A much larger proportion of Marines qualify as expert under the 350-point system than previous system (42 percent compared to 28 percent previously). This is likely because the BCRM gives Marines the opportunity to recover from a poor performance during the FRM course. For instance, a Marine who scores 205 during FRM would previously receive a qualification of marksman, but can realistically gain enough points during the BCRM course to achieve an expert qualification.

I converted the rifle scores of the three different scoring systems to the common rating used to compute composite scores (see Table 8). The promotion manual (USMC, 2012, p. 2–30) lists the conversion charts for the 65-point and 250-point scales. To convert the 350-point scale to ratings, I used the same method as the Enlisted Promotion branch within Manpower Management Division. That method uses the minimum rating of each qualification level from the old system and distributes the remainder of the ratings evenly throughout the qualification level. For instance, the minimum rating for expert qualification is 4.6, which applies to scores 40–44, 220–224, and 305–311 for the 65-point, 250-point, and 350-point scales, respectively. A higher number of experts under the 350-point scale also mean that the ratings are higher compared to the 250-point scale.

The average ratings from the 250-point and 350-point scales are 39.7 and 42.9, respectively.

Table 8. Rifle Scores Converted to a Single Rating Scale for Composite Score. Source: Lane Beindorf (personal communication, December 16, 2016).

65 scale	250 scale	350 scale	Points	Classification	Rationale (350 scale)
57-65	240-250	336-350	5	Expert	Starting point 96% of 350pts = 336pts = 5.0
53-56	235-239	328-335	4.9		
49-52	230-234	320-327	4.8		4.9 – 4.6 broken into increments of 7- or 8-pts
45-48	225-229	312-319	4.7		
40-44	220-224	305-311	4.6		
38-39	215-219	292-304	4.4	Sharpshooter	Broken into increments of 13 & 12 pts
35-37	210-214	280-291	4.2		
33-34	205-209	272-279	3.8	Marksman	Broken into increments of 7 or 8 pts
30-32	200-204	264-271	3.6		
28-29	195-199	257-263	3.4		
25-27	190-194	250-256	0		
0-24	0-189	0-249	0	Unqualified	

4. Time in Grade and Time in Service

I generate time in grade and time in service variables by subtracting the promotion date (for TIG) and armed forces active duty base date (for TIS) from the TFDW snapshot date and then dividing by 30.417 to generate months at snapshot date. I change impractical values, such as TIS beyond 96 months (8 years is the service limitation for a corporal [USMC, 2010]) or negative values, to missing values.

5. Personal Awards

I create several award categories in order to control for performance that may not be included in other performance measures. AI use dummy variables to indicate whether the Marine received an award during the snapshot period. Table 28 in Appendix A lists the awards associated with each award category.

6. Occupational Variables

I create technical and nontechnical PMOS dichotomous variables to allow for separate analysis between technical and nontechnical job fields. I code a PMOS as technical if the PMOS requires an Armed Services Vocational Aptitude Battery general technical, mechanical maintenance, or electronics repair score of 105 or greater. I code all remaining PMOSs with less stringent or no requirements as nontechnical. Table 29 in Appendix A lists the PMOSs included in either category.

To compare subsamples across different types of units, I create dummy variables for unit type categories. I categorize the unit types by level of command and mission type. For instance, I group all infantry, tank, artillery, and combat engineer battalions into *Unit_Ground* and all Marine Air Wing component squadrons into *Unit_Air*. Furthermore, I separate units that are higher than battalion and squadron level that perform in a headquarters capacity such as Regiments and Groups into ground, air, and logistics categories. I code the unit types using the variable named *Present_RUC*. The variable *Unit_Nontrad* is the catch-all for all unit types that are not easily categorized or that represent a relatively small number of Marines. I include the *Present_RUC* list associated with each unit type in Table 30 in Appendix A.

C. RELIABILITY

Reliability estimates determine if the proficiency and conduct marks are stable performance measures over time and if marks are consistent between raters. The following questions address reliability.

1. Stability

- Are proficiency and conduct marks consistent measures of performance over time?

To analyze stability, I use the standard deviation of marks, within technical and nontechnical PMOS categories, for 360,690 active duty Marines at the paygrade of E3 or E4 between 2006 and 2016. Standard deviations that are stable over time indicate to decision-makers that the information on performance levels is not changing from one

year to the next. Optimally, I would analyze the changes in scores for the same group between two periods when performance is unlikely to change. A change in scores could be an indication of systematic variance resulting from a poorly defined performance measure or insufficient rater training. Analysis of standard deviations over time will tell some of the same story, but with much less certainty because performance is not constant.

2. Interrater Reliability

- Do proficiency and conduct marks vary between rater?

The best way to construct accurate reliability estimates is to have two supervisors who observe the Marine's performance equally, complete an evaluation, and compare differences in scores. Unfortunately, the data do not support such a test. Consequently, I develop tests that attempt to hold all else constant except for the rater. To achieve this, I analyze the differences in average proficiency marks in grade within a PMOS held by first-term Marines randomly assigned to myriad different units. In theory, the aggregate performance and ability levels of the Marines should not vary greatly between units because they are randomly assigned. Therefore, we are able to observe any systematic differences in how marks are assigned between different units.

I estimate interrater reliability using regression analysis. I select just a few PMOSs in which Marines have similar billet responsibilities across a number of different unit types. Among the specialties that serve at many different unit types, I choose PMOSs of 0111 (Administrative Specialist), 0231 (Intelligence Specialist), 0621 (Field Radio Operator), and 3531 (Motor Vehicle Operator). For each regression, I restrict to one of the four PMOSs listed above and use only one unit type per regression so that the reference group is all other unit types. Model (1) is an ordinary least squares regression model.

(1) $Y_{it} = X_{it}\beta + C_{it}\gamma + \Sigma\delta FY_t + \mu_{it}$

where

- Y is the average proficiency mark in grade
- C is one of the seven unit types

- *X* is the set of control variables
- *FY* is the set of dummy variables for fiscal year

A significant effect of unit type on proficiency and conduct marks indicates different grading philosophies and thus low interrater reliability, all else being equal. However, results are not sufficient to determine the level of interrater reliability because the sample selection is non-random. Rather, the findings here warrant further examination of other PMOSs and unit type combinations to estimate systematic differences. Table 9 displays the variable definitions, and Table 10 contains the summary statistics.

Table 9. Variable Definitions.

VARIABLES	LABELS
Dependent Variables	
Proficiency_Grade	Average proficiency marks in grade
Conduct_Grade	Average conduct marks in grade
Demographic Variables	
Female	=1 if female, =0 if male
Single	=1 if single, =0 otherwise
Married	=1 if married, =0 otherwise
Marital_other	=1 if annul/divorce/sep/widow, =0 otherwise
White	=1 if white ethnicity, =0 otherwise
Hispanic	=1 if hispanic ethnicity, =0 otherwise
Black	=1 if black ethnicity, =0 otherwise
Asian	=1 if asian ethnicity, =0 otherwise
Ethnic_other	=1 if other ethnicity, =0 otherwise
Ethnic_declined	=1 if declined to respond, =0 otherwise
Performance Variables	
Rifle_rating	Rifle score converted to 5.0 scale
PFT_rating	PFT score converted to 5.0 scale
Composite_Duty_Bonus	=Total special duty bonus points
Composite_Educ_Bonus	=Total education bonus points
Composite_Recruiting_Bonus	=Total recruiting bonus points
High_Award_gain	=1 if gain high level award, =0 otherwise
Commend_Medal_gain	=1 if gain Commendatory Medal (any service), =0 otherwise
Achiev_Medal_gain	=1 if gain Achievement Medal (any service), =0 otherwise
Low_Award_gain	=1 if gain low level award, =0 otherwise
Combat_Action_gain	=1 if gain Combat Action Ribbon, =0 otherwise
LOA_gain	=1 if gain Letter of Appreciation, =0 otherwise
Volunteer_Medal_gain	=1 if gain Outstanding Volunt. Medal, =0 otherwise

VARIABLES	LABELS
Swim_basic	=1 if basic swim qualified, =0 otherwise
Swim_waiver	=1 if exempted from swim qualification, =0 otherwise
Swim_unq	=1 if unqualified, =0 otherwise
Swim_inst	=1 if swim instructor or instructor trainer qualified, =0 otherwise
Swim_advance	=1 if advanced swim qualified, =0 otherwise
Swim_inter	=1 if intermediate swim qualified, =0 otherwise
Swim_miss	=1 if swim qual data missing, =0 otherwise
Adverse_conduct	=1 if weight control, grade reduction, PFT/CFT fail, =0 otherwise
Occupational Variables	
E4	=1 if first time Marine held paygrade E4, =0 otherwise
E3	=1 if first time Marine held paygrade E3, =0 otherwise
TOS	=Months' time on station
TIG	=Months' time in grade
First_Dutystation	=1 if serving at first duty station, =0 otherwise
PMOS_0111	=1 if PMOS 0111, =0 otherwise
PMOS_0231	=1 if PMOS 0231, =0 otherwise
PMOS_0621	=1 if PMOS 0621, =0 otherwise
PMOS_3531	=1 if PMOS 3531, =0 otherwise
Unit_Ground	=1 if assigned to Bn level Ground unit, =0 otherwise
Unit_Air	=1 if assigned to Sqdn level Wing unit, =0 otherwise
Unit_Logistics	=1 if assigned to Bn level Logistics unit, =0 otherwise
HQ_Ground	=1 if assigned to Ground unit higher than Bn level, =0 otherwise
HQ_Air	=1 if assigned to Wing unit higher than Sqd/Bn level, =0 otherwise
HQ_Logistics	=1 if assigned to Logistics unit higher than Bn level, =0 otherwise
Unit_Nontrad	=1 if not assigned to Unit_XXX or HQ_XXX, =0 otherwise

Table 10. Summary Statistics—Reliability Analysis.

VARIABLES	N	mean	sd	min	max
Dependent Variables					
Proficiency_Grade	1561024	44.1251	1.6070	1	50
Conduct_Grade	1561015	44.0715	1.7391	1	50
Demographic Variables					
Female	1725022	0.0723	0.2589	0	1
Single	1725021	0.6433	0.4790	0	1
Married	1725021	0.3442	0.4751	0	1
Marital_other	1725021	0.0125	0.1112	0	1
White	1725022	0.5270	0.4993	0	1
Hispanic	1725022	0.1379	0.3448	0	1
Black	1725022	0.0639	0.2446	0	1
Asian	1725022	0.0250	0.1562	0	1
Ethnic_other	1725022	0.2461	0.4307	0	1
Ethnic_declined	1725022	0.2095	0.4070	0	1
Performance Variables					
Rifle_rating	1721893	41.7338	7.0310	0	50
PFT_rating	1718622	44.9441	6.3820	0	50
Composite_Duty_Bonus	1725022	1.5591	12.3887	0	100
Composite_Educ_Bonus	1725022	34.4110	43.2862	0	100
Composite_Recruiting_Bonus	1725022	0.6549	6.0703	0	100
High_Award_gain	1725022	0.0088	0.0934	0	1
Commend_Medal_gain	1725022	0.0017	0.0411	0	1
Achiev_Medal_gain	1725022	0.0533	0.2245	0	1
Low_Award_gain	1725022	0.2456	0.4304	0	1
Combat_Action_gain	1725022	0.0851	0.2790	0	1
LOA_gain	1725022	0.1697	0.3753	0	1
Volunteer_Medal_gain	1725022	0.0004	0.0211	0	1
Swim_basic	1713233	0.5691	0.4952	0	1
Swim_waiver	1713233	0.0126	0.1116	0	1
Swim_unq	1713233	0.0069	0.0828	0	1
Swim_inst	1713233	0.0017	0.0415	0	1
Swim_advance	1713233	0.0366	0.1878	0	1
Swim_inter	1713233	0.3731	0.4836	0	1
Swim_miss	1725022	0.0068	0.0824	0	1
Adverse_conduct	1725022	0.0482	0.2143	0	1
Occupational Variables					
E4	1725022	0.4480	0.4973	0	1
E3	1725022	0.5520	0.4973	0	1
TOS	1725017	18.1220	11.6835	0	60
TIG	1724837	12.2373	8.4907	0	60
First_Dutystation	1725022	0.8487	0.3584	0	1
For simplicity, this table does not display variables for PMOS and unit type.					

Table 11 and Table 12 contain the descriptive statistics for proficiency and conduct marks, respectively. The descriptive statistics show whether the difference in means between groups is significant or not. In both tables, I include the ratio of E3 to E4. The descriptive statistics with a high or low ratio are likely influenced by differences in the mean between paygrades rather than unit type. That is, a statistically significant difference in means when the ratio is close to 1.0 means that the level of significance is expected in the regression results as well.

Table 11. Descriptive Statistics for Average Proficiency Marks between Unit Type.

Variable	N_unit	N_total	Unit=1	Unit=0	T-stat	Ratio E3:E4
PMOS = 0111(Administrative Specialist)						
Unit_Ground***	2068	31918	44.085	44.787	17.230	1.63
Unit_Air***	2592	31918	44.616	44.757	3.852	1.05
Unit_Logistics**	781	31918	44.592	44.750	2.455	1.23
HQ_Ground	2830	31918	44.753	44.746	-0.213	1.13
HQ_Air	3305	31918	44.761	44.745	-0.515	1.17
HQ_Logistics***	3363	31918	44.431	44.780	10.891	1.52
Unit_Nontrad***	20411	31918	44.884	44.558	-17.001	1.05
PMOS = 0231 (Intelligence Specialist)						
Unit_Ground***	4304	16328	44.175	44.330	5.133	0.84
Unit_Air***	1917	16328	44.641	44.251	-9.362	0.75
Unit_Logistics	216	16328	44.382	44.291	-0.781	0.83
HQ_Ground***	5646	16328	44.219	44.326	3.826	0.83
HQ_Air***	1649	16328	44.434	44.278	-3.469	0.76
HQ_Logistics	378	16328	44.371	44.291	-0.880	0.74
Unit_Nontrad	3897	16328	44.285	44.295	0.324	1.25
PMOS = 0621 (Field Radio Operator)						
Unit_Ground**	28553	57998	43.857	43.889	2.415	0.72
Unit_Air**	4908	57998	43.920	43.871	-1.980	0.58
Unit_Logistics	3935	57998	43.843	43.877	1.257	0.65
HQ_Ground***	8435	57998	43.759	43.893	6.942	0.66
HQ_Air	38	57998	43.583	43.875	1.119	0.41
HQ_Logistics**	2042	57998	43.790	43.878	2.340	0.79
Unit_Nontrad***	17457	57998	43.968	43.842	-8.510	0.76
PMOS = 3531 (Motor Vehicle Operator)						
Unit_Ground	19451	84816	43.899	43.913	1.040	0.98
Unit_Air	8546	84816	43.899	43.911	0.652	0.70
Unit_Logistics***	20135	84816	43.833	43.931	7.600	0.99
HQ_Ground***	14140	84816	43.851	43.920	4.713	0.91
HQ_Air	173	84816	43.876	43.910	0.282	0.52
HQ_Logistics***	3095	84816	43.797	43.914	3.976	1.01
Unit_Nontrad***	27904	84816	44.018	43.864	-13.462	1.20

***Indicates statistical significance at 1 percent level

** Indicates statistical significance at 5 percent level

* Indicates statistical significance at 10 percent level

Table 12. Descriptive Statistics for Average Conduct Marks between Unit Type.

Variable	N_unit	N_total	Unit=1	Unit=0	T-stat	Ratio E3:E4
PMOS = 0111(Administrative Specialist)						
Unit_Ground***	2068	31917	44.216	44.774	12.908	1.63
Unit_Air**	2592	31917	44.661	44.748	2.260	1.05
Unit_Logistics*	781	31917	44.612	44.745	1.957	1.23
HQ_Ground	2830	31917	44.740	44.742	0.047	1.13
HQ_Air	3305	31917	44.734	44.743	0.262	1.17
HQ_Logistics***	3363	31917	44.405	44.778	10.973	1.52
Unit_Nontrad***	20411	31917	44.868	44.570	-14.626	1.05
PMOS = 0231 (Intelligence Specialist)						
Unit_Ground***	4304	16328	44.154	44.329	5.442	0.84
Unit_Air***	1917	16328	44.666	44.241	-9.582	0.75
Unit_Logistics	216	16328	44.299	44.287	-0.099	0.83
HQ_Ground***	5646	16328	44.178	44.337	5.369	0.83
HQ_Air***	1649	16328	44.494	44.266	-4.771	0.76
HQ_Logistics	378	16328	44.350	44.286	-0.668	0.74
Unit_Nontrad	3897	16328	44.311	44.280	-0.919	1.25
PMOS = 0621 (Field Radio Operator)						
Unit_Ground***	28553	57996	43.701	43.757	3.817	0.72
Unit_Air	4908	57996	43.765	43.730	-1.274	0.58
Unit_Logistics**	3935	57996	43.668	43.737	2.234	0.65
HQ_Ground***	8435	57996	43.649	43.745	4.431	0.66
HQ_Air	38	57996	43.639	43.733	0.319	0.41
HQ_Logistics	2042	57996	43.712	43.733	0.500	0.79
Unit_Nontrad***	17457	57996	43.835	43.696	-8.360	0.76
PMOS = 3531 (Motor Vehicle Operator)						
Unit_Ground***	19451	84816	43.789	43.826	2.578	0.98
Unit_Air	8546	84816	43.834	43.816	-0.876	0.70
Unit_Logistics***	20135	84816	43.730	43.842	7.820	0.99
HQ_Ground***	14140	84816	43.773	43.826	3.266	0.91
HQ_Air	173	84816	43.938	43.818	-0.891	0.52
HQ_Logistics***	3095	84816	43.703	43.822	3.629	1.01
Unit_Nontrad***	27904	84816	43.932	43.770	-12.673	1.20

***Indicates statistical significance at 1 percent level

** Indicates statistical significance at 5 percent level

* Indicates statistical significance at 10 percent level

D. VALIDITY

I use quantitative analysis to answer the questions about construct validity and predictive validity. I use factor analysis to answer the question related to the construct validity of the composite score, which reveals the underlying correlations between performance variables. Once I discover the underlying correlations, I use the latent variables to estimate predictive validity.

1. Construct Validity

Construct validity takes all the parts of one performance measure and tests that they measure what they are intended to measure. For this analysis, the composite score is the one performance measure, and its parts are 10 performance-related variables including proficiency and conduct marks. Proficiency and conduct marks have subparts as well, but the data are not available to conduct a separate analysis. Therefore, I conduct factor analysis on the composite score to determine the relative importance of proficiency and conduct marks among the other composite score components and whether the components' relative importance corresponds with their weighted contribution to the composite score. Additionally, factor analysis reveals the latent variables or groups of variables with underlying correlations, among the 10 performance-related components. The following research question addresses construct validity:

- Of the 10 performance-related measures comprising the composite score, which variables provide the most information on the Marine's performance level?

Proficiency and conduct marks should provide the most information on a Marine's performance because the marks are the most heavily weighted components of the composite score. Moreover, I expect proficiency and conduct marks to be closely associated despite the intended distinction between the two marks. Proficiency should measure the Marine's ability and effectiveness in performing his or her primary duty and should correlate with physical fitness and education. Conduct should measure the Marine's conformance to the organizational norms, regulations, and standards, and should correlate with physical fitness as well.

a. Composite Score Data

The dataset is restricted to Marines who hold the paygrade E3 or E4, were not previously promoted or reduced to or from E3 or E4, and do not have missing values for any of the composite score components. The data consist of 979,353 observations and 240,864 distinct individuals. Among the promotion-eligible, 368,921 observations and 170,616 distinct individuals are at the paygrade E3, and 189,938 observations and 103,087 distinct individuals are E4.

I transform composite score components to the point scale and weights used to calculate the actual composite score. I average rifle and PFT and CFT ratings, then multiply the average by 100. I multiple proficiency and conduct marks by 100, months' time in grade by 5, and months' time in service by 2. Special duty, self-education, and command recruiting bonus points do not require conversion prior to composite score computation. Table 13 provides the variable definitions.

Table 13. Composite Score Variable Definitions.

Variable	Definition
Composite_Rifle	=Rifle rating x 33.3
Composite_PFT	=PFT rating x 33.3
Composite_CFT	=CFT rating x 33.3
Composite_Pro	=Average proficiency marks in grade x 100
Composite_Con	=Average conduct marks in grade x 100
Composite_TIG	=TIG (months) x 5 if eligible
Composite_TIS	=TIS (months) x 2 if eligible
Composite_Duty_Bonus	=Total special duty bonus points
Composite_Educ_Bonus	=Total education bonus points
Composite_Recruiting_Bonus	=Total recruiting bonus points

b. Eligibility

Within each paygrade, I analyze only the promotion-eligible. I create a dummy variable for promotion eligibility that equals one if the Marine is eligible for a composite score and does not have a promotion restriction including PFT or CFT failure, is not assigned to weight control, and is not flagged as "not recommended for promotion." I use

only the promotion-eligible to reveal the important factors affecting composite score among those competing for promotion. Table 14 lists the summary statistics for each component, and Table 15 displays the difference in means (*t*-statistics) between E3 and E4 and technical and nontechnical PMOS subsamples.

Table 14. Promotion Eligible Composite Score Summary Statistics by Paygrade.

VARIABLES	N	mean	sd	min	max
E4					
Composite_Rifle	189,938	145.8	21.5	0	167
Composite_PFT	189,938	154.0	8.0	100	167
Composite_CFT	189,938	157.5	7.5	0	167
Composite_Pro	189,938	447.1	13.4	200	500
Composite_Con	189,938	446.8	13.9	180	500
Composite_TIG	189,938	101.1	32.1	60	300
Composite_TIS	189,938	96.7	20.1	48	192
Composite_Educ_Bonus	189,938	62.1	42.2	0	100
Composite_Duty_Bonus	189,938	1.1	10.6	0	100
Composite_Recruiting_Bonus	189,938	0.8	6.6	0	100
E3					
Composite_Rifle	368,921	141.1	23.5	0	167
Composite_PFT	368,921	152.6	8.4	100	167
Composite_CFT	368,921	157.0	7.4	0	167
Composite_Pro	368,921	435.5	12.8	180	500
Composite_Con	368,921	435.3	13.6	100	500
Composite_TIG	368,921	83.5	34.4	40	295
Composite_TIS	368,921	56.7	15.1	24	192
Composite_Educ_Bonus	368,921	71.4	36.3	0	100
Composite_Duty_Bonus	368,921	2.6	16.0	0	100
Composite_Recruiting_Bonus	368,921	0.1	1.9	0	100

Table 15. Composite Score Descriptive Statistics *t*-Test Results.

Variable	E3	E4	sig level	E3			E4		
				Tech	Nontech	sig level	Tech	Nontech	sig level
Composite_Rifle	141.89	144.60	***	142.82	141.44	***	145.23	144.23	***
Composite_PFT	149.84	152.13	***	148.91	150.29	***	151.26	152.66	***
Composite_CFT	156.64	157.77	***	156.32	156.79	***	157.34	158.03	***
Composite_Pro	434.75	445.59	***	435.46	434.40	***	445.12	445.88	***
Composite_Con	434.38	445.35	***	434.96	434.10	***	444.90	445.62	***
Composite_TIG	67.33	67.60	***	68.32	66.85	***	66.90	68.02	***
Composite_TIS	50.49	85.79	***	51.13	50.19	***	85.08	86.22	***
Composite_Educ_Bonus	50.71	31.31	***	52.44	49.87	***	32.30	30.71	***
Composite_Duty_Bonus	2.59	0.99	***	0.09	3.80	***	0.70	1.17	***
Composite_Recruiting_Bonus	0.09	0.44	***	0.09	0.09		0.18	0.60	***
*** indicates significance at 1 percent									

c. Model Selection

Among the promotion-eligible, the most significant differences in factor loadings are between paygrade and technical versus nontechnical PMOS. The descriptive statistics in Table 15 confirm these differences. I explore several other samples restricted or unrestricted by paygrade and OccFld, which all show comparable results. Next, I choose the number of factors to retain based on four criteria: (1) eigenvalue is greater than one, (2) the number of retained factors account for more than 80 percent of the total variance, (3) the scree plot flattens out at the first discarded factor (Rencher & Christensen, 2012), and (4) the factor loadings are interpretable. If the first three criteria do not reveal a clear answer, I decide on the number factors based on the interpretability of the rotated factor loadings. Table 16 shows the number of factors to retain based on each criterion.

Table 16. Number of Factors to Retain by Sample and Criterion.

Sample	Eigenvalue	Proportion of variance	Scree plot	Interpret-ability
E3-technical	2	2	2	2
E4-technical	2	2	3	3
E3-nontechnical	2	2	3	3
E4-nontechnical	2	2	3	3

Factor rotation—the next step of factor analysis—simplifies the interpretation of the factor loadings in which the variables are grouped into unique factors (Rencher & Christensen, 2012). Orthogonal and oblique rotations are the two main options for factor rotation. The primary difference between the two is that oblique rotation allows for the factors to be correlated. If an oblique rotation shows correlation greater than .32, then oblique is the preferred rotation method (Brown, 2009). Each model has at least one factor pair with correlation greater than .32; therefore, I choose oblique rotation for all models.

I retain three factors based on the interpretation of the factor loadings and scree plot. The scree plot shows an obvious elbow following the third factor for each of the subsamples (Figure 6). Comparisons of the two and three factor models reveal a much simpler structure with three factors, especially for the E4 samples (see Table 31 in Appendix B). Unfortunately, the number of factors to retain is not conclusive, which introduces some level of subjectivity into the process (Rencher & Christensen, 2012, p. 455). Nonetheless, I choose three factors for each subsample because interpretation is simpler.

Table 17. Eigenvalue and Proportion of Variance.

Factor	Eigenvalue	Difference	Proportion	Cumulative
E3 Technical PMOS				
Factor1	**2.07083**	0.55502	0.5866	0.5866
Factor2	**1.51581**	1.078	0.4293	**1.0159**
Factor3	0.4378	0.39809	0.124	1.1399
E3 Nontechnical PMOS				
Factor1	**2.0201**	0.59802	0.5743	0.5743
Factor2	**1.42208**	0.84064	0.4043	**0.9786**
Factor3	0.58144	0.53292	0.1653	1.1439
E4 Technical PMOS				
Factor1	**1.87728**	0.74373	0.6453	0.6453
Factor2	**1.13354**	0.64509	0.3896	**1.0349**
Factor3	0.48846	0.4411	0.1679	1.2028
E4 Nontechnical PMOS				
Factor1	**1.92159**	0.59942	0.608	0.608
Factor2	**1.32217**	0.84057	0.4183	**1.0263**
Factor3	0.4816	0.42002	0.1524	1.1787

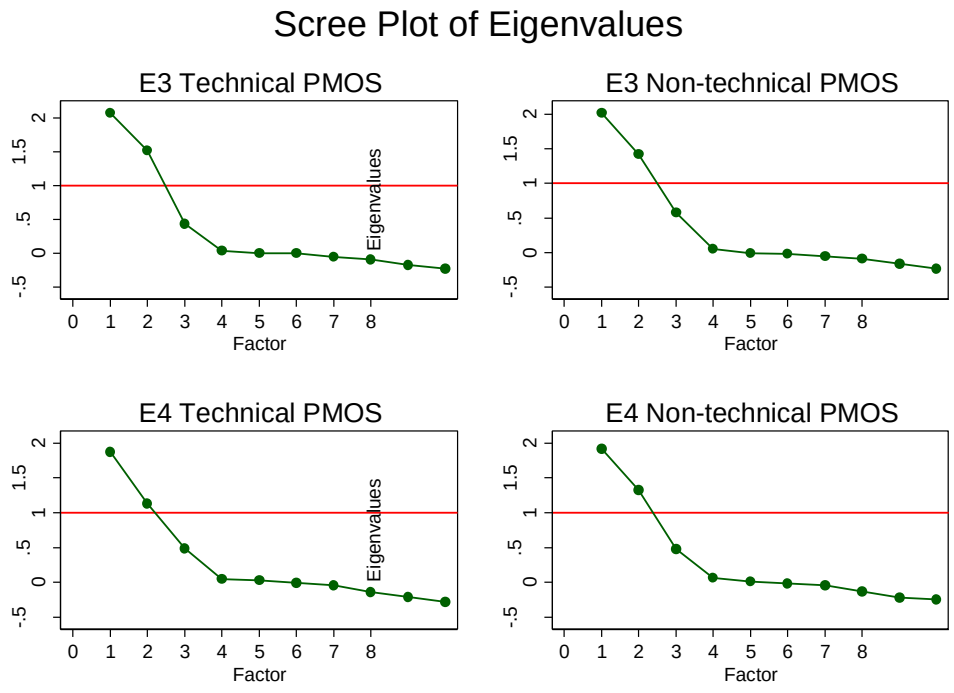

Figure 6. Scree Plot for Promotion-Eligible E3 and E4 by PMOS.

d. Standardized Latent Variables

I create common factors of each of the three retained factors, also known as latent variables, to use as explanatory variables in the models for predictive validity. I use the statistical software Stata "predict" command to create the standardized latent variables from the rotated factors. I choose to use the common factors instead of the individual performance variables to reduce the collinearity between variables and to improve the interpretability of the multivariate regression results. In theory, the common factor has a mean of zero and a standard deviation of one, although the estimation method rarely yields such exact results (StataCorp, 2013).

2. Predictive Validity

The performance-related elements that make up the composite score are used to provide information on a Marine's potential to perform at the next higher grade. Therefore, the Marine's performance scores should not be systematically different from his or her performance scores at the next higher grade. The following research question addresses predictive validity:

- Do proficiency and conduct marks predict future performance as indicated by fitness report scores?

I use the three latent variables from factor analysis to analyze the predictive validity of proficiency and conduct marks. I use only the E4 technical and nontechnical PMOS subsamples for analysis. I compare an individual's average scores in each composite score performance variable as a corporal with their fitness report scores for up to three years following promotion to sergeant. For Model (2), the dependent variables are the average reporting senior relative value (RSRV) and reviewing officer cumulative value (ROCV) weighted by the months of observation during the reporting period. The RSRV is a value ranging from 80 to 100, and the ROCV is the number of marks the Marine is from the RO's average mark. A zero ROCV is the RO's average, a negative value is below average, and a positive value is above average. Model (3) uses dichotomous variables that represent a ROCV in the top third or above 93.33 and a ROCV that is above average or greater than zero.

Model (2) is an ordinary least squares regression model that estimates the predictive power of the latent performance variables on RSRV or ROCV average. I run separate models for the technical and nontechnical PMOS subsamples.

(2) $Y_i = X_i\beta + \mu_i$

where

- Y is the RSRV or ROCV weighted average
- X is the set of latent performance variables

Model (3) is a probit model that estimates the partial effect of the latent performance variables on the probability of having an RSRV in the top third or an above average ROCV. I use Stata's "dprobit" command to estimate the partial effect. Again, I run separate models for the technical and nontechnical PMOS subsamples.

(3) $P(y = 1 | X) = G(X_i\beta + \mu_i)$

where

- y is the dichotomous variable, top third RSRV, or above average ROCV
- X is the set of latent performance variables

I use robust standard errors clustered by individual to correct for heteroscedasticity because of an unbalanced sample. The latent independent performance variables from factor analysis included multiple observations per Marine. If I were to use only the last observation prior to promotion to sergeant, then the latent performance variable would lose standardization. Therefore, Model (2) and Model (3) compare multiple observations of a Marine at the rank of corporal to a single dependent variable value of the same Marine at the rank of sergeant.

E. ACCURACY

Accurate marks are representative of the Marine's true performance level, whereas inaccurate marks misinform decision-makers about a Marine's true performance. Central tendency error, rater leniency, and halo error are the causes of inaccuracy. I analyze data trends to determine levels of accuracy. Admittedly, the process is somewhat

subjective because it is not clear how much inflation is too much inflation, or how much central tendency is too much central tendency. The analysis does reveal, however, the presence of potentially harmful errors relative to the effectiveness of proficiency and conduct marks. The following research questions relate to issues of accuracy:

- Do proficiency and conduct marks differentiate between levels of performance?

This question relates to the rater's tendency to assign average marks or to avoid using the extreme ends of the scale. Analysis of the standard deviation indicates how much differentiation exists among the marks between individuals. Factor analysis also indicates how much of the total variance in performance characteristics is explained by the variance in proficiency and conduct.

- Are proficiency and conduct marks subject to rater leniency?

Rater leniency, or inflation, means that raters assign higher scores than is warranted by an individual's performance. The extent of rater leniency is determined by comparing actual distribution of marks with the intended distribution.

- Are proficiency and conduct marks distinct measures of performance?

This question refers to the halo effect, or the rater's tendency to assign similar scores to several performance measures. It also refers to the rater's tendency to assign marks based on previous marks or the occurrence of logical error when the rater has difficulty understanding the differences between two similar measures. A high correlation between proficiency and conduct marks would suggest a possible halo error, though it may also be attributed to the systematic causes of a poor criterion construct.

F. PRACTICALITY

To determine practicality, I make a qualitative assessment based mainly on the observability and interpretability of the performance attributes listed for proficiency and conduct in the IRAM. Also, I make a determination of the marks' usability. I incorporate that assessment into the conclusions and recommendations, keeping in mind the explicit and implicit costs related to making changes to the performance evaluation system.

THIS PAGE INTENTIONALLY LEFT BLANK

V. RESULTS AND ANALYSIS

A. RELIABILITY

There is no evidence to suggest that proficiency and conduct marks are instable, only that they are potentially becoming less informative. Interrater reliability estimates show that proficiency and conduct marks are different between units within a PMOS, all else being equal. The most significant effect is among administrative specialists who, because of low interrater reliability, may be receiving different proficiency and conduct marks for equal performance.

1. Stability Estimates

Stability estimates answer the following research question:

- Are proficiency and conduct marks consistent measures of performance over time?

Proficiency and conduct marks appear to be stable, and there is some evidence to suggest that both marks respond appropriately to performance changes. Figure 7 depicts the standard deviation trend across fiscal year for paygrades E3 and E4 in a technical and nontechnical PMOS. Proficiency marks are relatively more stable for all groups. Conduct marks, however, are declining at an annualized rate of 5.54 percent for lance corporals and 2.19 percent for corporals in terms of standard deviation. A declining standard deviation means that marks are potentially becoming less informative.

The changes in average standard deviation among the subsamples appear to be more volatile from 2006 to 2010 than in recent years. That may be a result of the buildup of forces during the wars in Iraq and Afghanistan and a more diverse population of Marines in terms of quality. This indicates that proficiency and conduct are sensitive to changes in performance, which further reinforces that the marks are stable, rather than just appearing to be stable. Lastly, subjective evaluations, like proficiency and conduct marks, are more stable when the marks pass through multiple evaluators (Lazear & Gibbs, 2015).

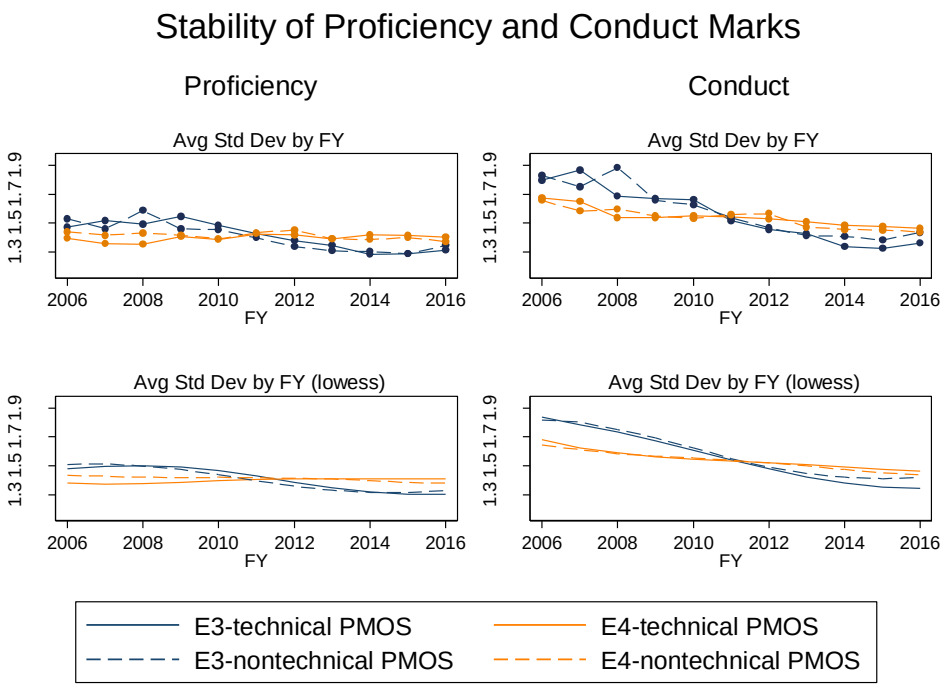

Figure 7. Stability of Proficiency and Conduct Marks by Paygrade and PMOS Category.

2. Interrater Reliability

There is weak evidence to support the claim that proficiency and conduct marks are inconsistent between raters. I analyze the effect of unit for each PMOS subsample on the average proficiency marks in grade for all Marines in that particular unit. The interpretation of the effect of unit type on proficiency and conduct marks is in reference to all other unit types. I use Model (1) to answer the following research question and display the results in Table 18:

- Do proficiency and conduct marks vary between rater?

Table 18. Effect of Unit Type on Proficiency Marks by PMOS.

VARIABLES	PMOS_0111	PMOS_0231	PMOS_0621	PMOS_3531
Dependent Variable = Proficiency_Grade				
Unit_Ground	-0.4277***	-0.2173***	-0.0694***	-0.0199*
	(0.0346)	(0.0252)	(0.0122)	(0.0114)
Unit_Air	-0.1677***	0.2924***	0.0682***	-0.0819***
	(0.0336)	(0.0334)	(0.0213)	(0.0158)
Unit_Logistics	-0.1810***	-0.0536	-0.0230	-0.0643***
	(0.0529)	(0.0909)	(0.0235)	(0.0109)
HQ_Ground	0.0344	-0.0586**	-0.0996***	-0.0600***
	(0.0289)	(0.0234)	(0.0171)	(0.0128)
HQ_Air	-0.0459*	0.0720	-0.2058	-0.0495
	(0.0264)	(0.0438)	(0.2347)	(0.1274)
HQ_Logistics	-0.2107***	0.0610	0.0512	-0.1618***
	(0.0271)	(0.0721)	(0.0324)	(0.0261)
Unit_Nontrad	0.2429***	0.0951***	0.1142***	0.1705***
	(0.0166)	(0.0264)	(0.0134)	(0.0103)
Dependent Variable = Conduct_Grade				
Unit_Ground	-0.3112***	-0.2080***	-0.0835***	-0.0353***
	(0.0348)	(0.0268)	(0.0140)	(0.0130)
Unit_Air	-0.1320***	0.3038***	0.0528**	-0.0476***
	(0.0339)	(0.0378)	(0.0239)	(0.0181)
Unit_Logistics	-0.1546***	-0.1646	-0.0630**	-0.0756***
	(0.0533)	(0.1007)	(0.0261)	(0.0125)
HQ_Ground	0.0153	-0.0771***	-0.0681***	-0.0403***
	(0.0312)	(0.0251)	(0.0191)	(0.0145)
HQ_Air	-0.0606**	0.1284***	0.1069	0.1531
	(0.0286)	(0.0420)	(0.2330)	(0.1234)
HQ_Logistics	-0.2107***	0.0524	0.0959***	-0.1415***
	(0.0293)	(0.0769)	(0.0361)	(0.0292)
Unit_Nontrad	0.2156***	0.0840***	0.1208***	0.1600***
	(0.0176)	(0.0305)	(0.0156)	(0.0117)
Observations	31,728	16,104	57,219	83,773

This table displays coefficient estimates from 28 separate regressions. Each regression varies only by PMOS and unit type.

This table excludes the constant and control variables.

Dependent variables are scale x 10.

Robust standard errors are in parentheses.

*** p < .01, ** p < .05, * p < .1

Most of the coefficients are statistically significant, though probably not practically significant. The effect of being assigned to a ground unit or a nontraditional unit is the strongest among the coefficients and is likely to be practically significant as well. For example, administrative specialists (PMOS 0111) assigned to a ground unit are expected to receive proficiency marks that are 0.04277 points, or 25 percent of a standard deviation, lower on average than their peers. The practical significance is low because the average effect is the difference between a 4.5 and a 4.4572. The results for conduct marks are similar to proficiency marks for all samples.

B. VALIDITY

Factor analysis reveals the underlying construct of the composite score and shows that proficiency and conduct marks together, compared to all other performance elements in the composite score, is the most important factor for corporals. Predictive validity results from the multivariate regression model show that the underlying performance element of proficiency and conduct is a significant predictor of future performance.

1. Construct Validity

Factor analysis reveals three underlying performance elements within the composite score, which I call *person–organization fit*, *physical fitness*, and *human capital*. Three of the 10 performance scores fail to load on any of the factors and thus provide little information about the Marine's performance. I use factor analysis to answer the following research question:

- Which composite score variables provide the most information on the Marine's performance level?

Table 19 and Table 20 show the factor loadings along with how much variance each factor accounts for, as well as the variable's uniqueness. The variable's uniqueness is the amount that the variable does not have not in common with other variables (Rencher & Christensen, 2012).

Table 19. Factor Loadings for E3 Sample.

Variable	Technical PMOS				Nontechnical PMOS			
	Factor1	Factor2	Factor3	Uniqueness	Factor1	Factor2	Factor3	Uniqueness
Composite_Rifle				0.97				0.9617
Composite_PFT			0.5764	0.6369			0.5608	0.6663
Composite_CFT			0.5951	0.6555			0.5799	0.6685
Composite_Pro		0.8204		0.2962		0.8348		0.2713
Composite_Con		0.8432		0.3198		0.8447		0.2902
Composite_TIG	0.9252			0.141	0.927			0.141
Composite_TIS	0.9202			0.135	0.9199			0.1382
Composite_Educ_Bonus	0.3744			0.824				0.8802
Composite_Duty_Bonus				0.9983				0.9603
Composite_Recruiting_Bonus				0.9989				0.9987
Variance Accounted For	0.5866	0.4293	0.124		0.5743	0.4043	0.1653	
Blanks represent absolute value loadings < .3.								

Table 20. Factor Loadings for E4 Sample.

Variable	Technical PMOS				Nontechnical PMOS			
	Factor1	Factor2	Factor3	Uniqueness	Factor1	Factor2	Factor3	Uniqueness
Composite_Rifle				0.9877				0.9829
Composite_PFT		0.6104		0.6059			0.5861	0.6408
Composite_CFT		0.5973		0.6465			0.5948	0.6568
Composite_Pro	0.8779			0.2236	0.8822			0.2085
Composite_Con	0.8923			0.235	0.8983			0.221
Composite_TIG			0.7206	0.4916		0.7936		0.3862
Composite_TIS			0.6102	0.5345		0.7076		0.4434
Composite_Educ_Bonus			0.3237	0.8145		0.3218		0.8205
Composite_Duty_Bonus				0.9967				0.9971
Composite_Recruiting_Bonus				0.9647				0.9175
Variance Accounted For	0.6453	0.3896	0.1679		0.608	0.4183	0.1524	
Blanks represent absolute value loadings < .3.								

The factor loadings for E3 do not change between the technical and nontechnical PMOS samples. This indicates that although the mean scores are different between samples (Table 15), the variance is similar. In the E3 sample, there is more variation in

TIG, TIS, and education than in proficiency and conduct marks, unlike the E4 results. The factor loadings for the E4 technical and nontechnical PMOS subsamples are different. This suggests that variation in physical fitness scores and experience are different between Marines with a technical PMOS and Marines with a nontechnical PMOS. This may be because nontechnical specialties have higher PFT and CFT scores on average and thus less variation.

The factor loadings reveal the underlying performance elements within the composite score. Table 21 lists the factor labels and the associated performance variables. Proficiency and conduct appear to be a measure of person–organization fit or overall compatibility with the Marine Corps, one's assigned unit, and specialty. PFT and CFT are clearly a measure of physical fitness, and TIG, TIS, and self-education appear to measure the Marine's human capital gained through experience and education.

Table 21. Factor Labels and Associated Performance Measures.

Factor Label	Performance Measures
Person–Organization Fit	Proficiency, Conduct
Physical Fitness	PFT, CFT
Human Capital	TIG, TIS, Educ_Bonus
Independent Measures	Rifle, Duty_Bonus, Recruiting_Bonus

Factor analysis also reveals the independent elements: rifle score, special duty bonus, and recruiting bonus. Special duty and recruiting bonuses have such little variation that it is not surprising they did not associate with any other variables. Interestingly, though, rifle marksmanship did not associate with any other variables. That means that the individual characteristics that contribute to rifle marksmanship are not common to physical fitness, person–organization fit, or human capital. The findings are somewhat contrary to the findings of Chung et al. (2011), which identified a relationship between rifle marksmanship and aptitude, psychomotor skills, and non-cognitive aspects such as anxiety. The bottom line is that it is difficult to say what exactly rifle marksmanship represents about the Marine's performance.

Another interesting finding is that education is more closely associated with time spent in the Marine Corps and in grade, although just slightly, than with proficiency marks despite self-education being explicitly stated as a part of the consideration when assigning proficiency marks (USMC, 2000). That could be an indication that raters do not follow the guidance in the IRAM, or that education is not something raters feel to be part of the "whole Marine concept" as stated in the IRAM (USMC, 2000, p. 4–42).

2. Predictive Validity

I take the latent performance variables derived from factor analysis to analyze the extent to which they predict the Marine's fitness report scores as a sergeant. I use Model (2) and Model (3) to answer the following research question:

- Do proficiency and conduct marks predict future performance as indicated by fitness report scores?

Compared to the other performance variables in the model, proficiency and conduct marks are the most powerful predictor of future performance, as shown in Table 22 and Table 23.

Table 22. Predictive Validity Results for E4 Nontechnical PMOS.

Nontechnical PMOS				
VARIABLES	RSRV	ROCV	RSRV>93.33	ROCV>0
Person–Org Fit	1.4668***	0.3217***	0.0928***	0.1562***
	(0.0555)	(0.0125)	(0.0052)	(0.0072)
Physical Fitness	0.7347***	0.1863***	0.0366***	0.0721***
	(0.0660)	(0.0145)	(0.0063)	(0.0082)
Human Capital	-0.1816***	-0.0322***	-0.0105***	-0.0205***
	(0.0345)	(0.0077)	(0.0032)	(0.0044)
Constant	89.3501***	-0.2195***		
	(0.0462)	(0.0104)		
Observations	25,624	26,443	27,742	27,742
R-squared	0.1706	0.1738		
obs_P			0.183	0.452
Pseudo_R_sq			0.0733	0.0851
Coefficients for RSRV>93.33 and ROCV>0 are partial effects. Robust standard errors are in parentheses. *** p < .01, ** p < .05, * p < .1				

Table 23. Predictive Validity Results for E4 Technical PMOS.

Technical PMOS				
VARIABLES	RSRV	ROCV	RSRV>93.33	ROCV>0
Person–Org Fit	1.4227***	0.2992***	0.0797***	0.1383***
	(0.0635)	(0.0146)	(0.0055)	(0.0082)
Physical Fitness	0.8262***	0.2027***	0.0543***	0.0876***
	(0.0761)	(0.0165)	(0.0070)	(0.0096)
Human Capital	0.0446	-0.0025	0.0050	-0.0005
	(0.0457)	(0.0101)	(0.0047)	(0.0060)
Constant	89.2284***	-0.2639***		
	(0.0495)	(0.0112)		
Observations	16,047	17,249	18,945	18,945
R-squared	0.1546	0.1474		
obs_P			0.158	0.412
Pseudo_R_sq			0.0710	0.0669
Coefficients for RSRV>93.33 and ROCV>0 are partial effects. Robust standard errors are in parentheses. *** p < .01, ** p < .05, * p < .1				

The results for nontechnical PMOSs are slightly higher than for technical PMOSs. This indicates that performance scores as an E4 may be more valid for a Marine with a nontechnical PMOS. The mean scores are much lower than expected for the population of sergeants. *Constant* and *obs_P* show the mean scores of the dependent variable for each sample. By definition, the mean RSRV is 90 and the mean ROCV is zero. Approximately 33 percent of the population should have an RSRV greater than 93.33, and 50 percent should have a ROCV greater than 0. The lower than expected averages are likely a result of this study observing the performance scores of relatively junior sergeants. Thus, the magnitude of the effects is biased downward.

The importance of the latent performance variables coincide with the results of the factor analysis. Person–organization fit accounts for over 60 percent of the variance in the composite score and has the largest effect in the predictive models. For E4s in a nontechnical PMOS, a one standard deviation increase in person–organization fit is expected to increase the RSRV by 1.467 points, or 7.3 percent. Additionally, a one standard deviation increase in person–organization fit is expected to increase the probability that the Marine has an RSRV greater than 93.33 by 0.0928, or 50.7 percent. The results are similar, though slightly lower, for those with a technical PMOS.

C. ACCURACY

The quantitative results from stability tests, interrater reliability, factor analysis, and predictive validity provide additional information about the accuracy of proficiency and conduct marks. I use these results along with summary statistics to answer the following research questions. The first question relates to central tendency error:

- Do proficiency and conduct marks differentiate between levels of performance?

There is no evidence that proficiency and conduct marks fail to discriminate between different levels of performance. Factor analysis results suggest that proficiency and conduct are not susceptible to central tendency error. The variance in proficiency and conduct marks account for the majority of the variance in the composite score, which means that the marks provide at least some information of different performance levels.

Relatively speaking, proficiency and conduct provide more information on performance levels of E4s than any of the other composite score elements.

Both marks are approximately normally distributed, which means that about 68 percent of the sample is within one standard deviation of the mean (Keller, 2009, p. 111). That is, 68 percent of E3s receive marks approximately between 4.2 and 4.5 and the same percentage of E4s between 4.3 and 4.6. The separation of scores among the "average" performers may be sufficient to identify slightly above average performers from the truly average performers and likewise for slightly below average performers. Table 24 shows the summary statistics of proficiency and conduct marks for E3s and E4. Figure 8 and Figure 9 display the normal distributions of proficiency and conduct marks for E3s and E4s.

Table 24. Summary Statistics of Mean Proficiency and Conduct Marks in Grade.

VARIABLES	N	mean	sd	min	max
E3					
Proficiency_Grade	908462	43.6116	1.5657	1	50
Conduct_Grade	908454	43.5278	1.7503	1	50
E4					
Proficiency_Grade	702385	44.7433	1.5417	5	50
Conduct_Grade	702385	44.6891	1.6758	1	50

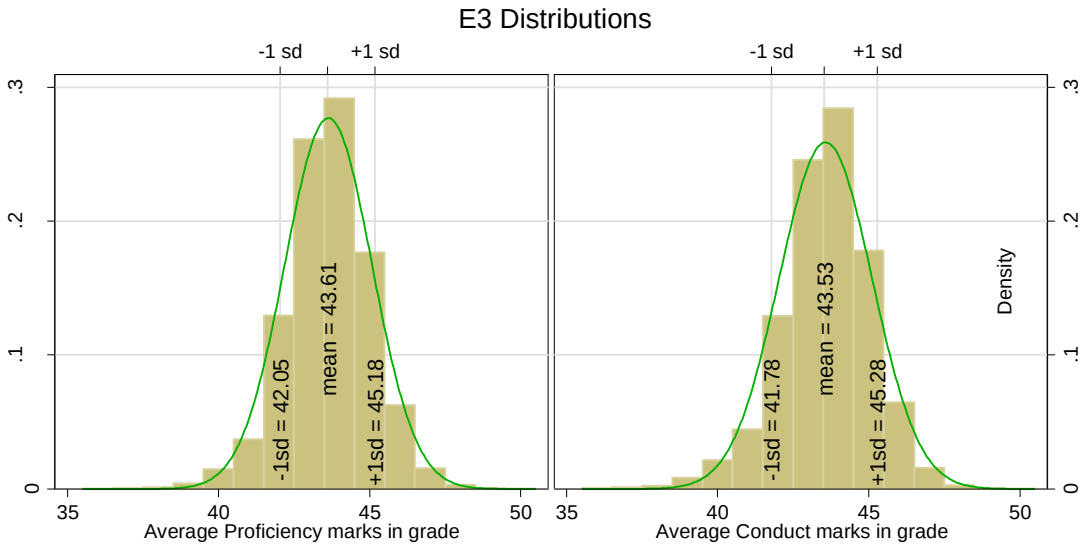

Figure 8. E3 Distributions of Average Proficiency and Conduct Marks in Grade.

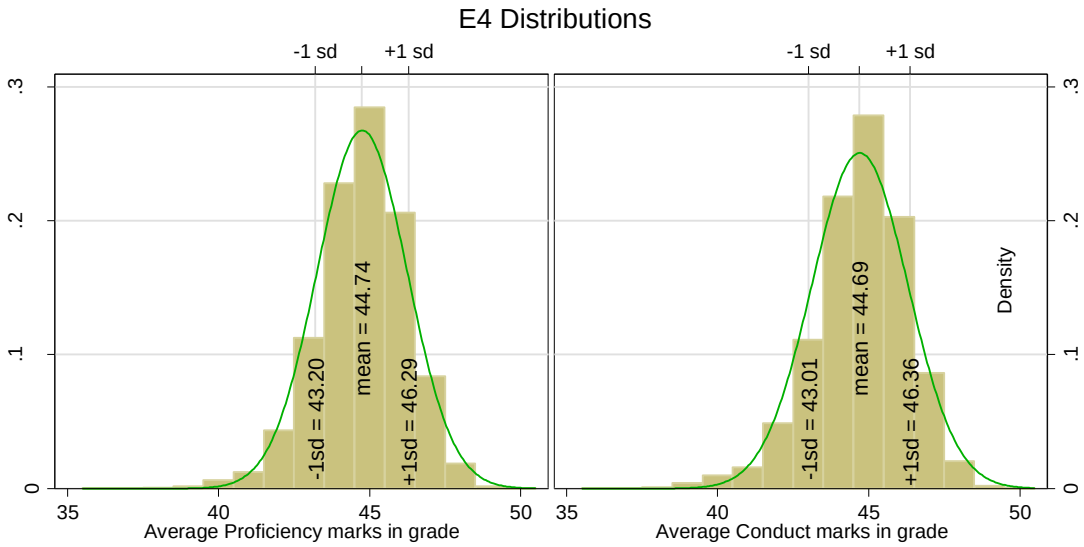

Figure 9. E4 Distributions of Average Proficiency and Conduct Marks in Grade.

The next question addresses the issue of inflation:

- Are proficiency and conduct marks subject to rater leniency?

The average proficiency and conduct marks are slightly higher than what is intended per the IRAM. According to the IRAM, average marks are between 4.0 and 4.4 (USMC, 2000). Therefore, the expected mean should be about 4.2. In 2016, the average proficiency mark is somewhere between 0.14 and 0.25 higher than the intended average of 4.2 (Table 25). The inflation is most likely due to the restriction of range at the lower end of the scale. The lowest conduct mark that can be assigned without supporting documentation is 4.0, indicating that the usable scale for the majority of the population is between 4.0 and 5.0. Although proficiency marks below 4.0 do not normally require supporting documentation, many may perceive the marks as "adverse" in nature because of the adverse nature of conduct marks below 4.0.

Inflation is not necessarily harmful to the promotion system that is based on relative performance within PMOS. On the other hand, inflation distorts the information on performance when considering Marines for competitive programs. For example, as previously discussed in Chapter II, a Marine must have a minimum of 4.4/4.4 proficiency/conduct marks to be considered for independent duty (USMC, 2001). Using this metric alone, an "average" administrative Marine would be more qualified to serve in an independent duty in terms of proficiency and conduct marks than an "average" tank Marine.

Table 25. OccFlds with Highest and Lowest Mean Proficiency and Conduct Marks in Grade in FY2016.

	Variable	Description	N	Mean Pro	Mean Con
Highest	OccFld_44	Legal Services	344	44.51	44.78
	OccFld_59	Aviation C2 Electronics Maint.	1447	44.38	44.13
	OccFld_01	Administrative	5748	44.30	44.32
Lowest	OccFld_03	Infantry	29689	43.41	43.38
	OccFld_61	Aircraft Maint.	6519	43.41	43.54
	OccFld_18	Tank and AAV	2089	43.40	43.50

Sample contains E3 and E4 only.
C2 = command and control; AAV = Amphibious Assault Vehicle

The next question addresses the halo effect, or the tendency for raters to anchor the mark of one performance measure to another or previous mark:

- Are proficiency and conduct marks distinct measures of performance?

There is strong evidence to support that proficiency and conduct marks are measuring much of the same performance. Proficiency and conduct marks are highly correlated. Using the entire sample of active duty Marines in the paygrades E1–E4 from 2006–2016, the correlation between proficiency and conduct marks is .84. Additionally, the results of a univariate regression show a highly collinear relationship between the two marks, as shown in Table 26. Most remarkably, a one point increase in proficiency mark is expected to increase the conduct mark by 0.95, on average.

Factor analysis provides additional evidence that the marks are not very distinct. Both marks have nearly identical weights on the factor to which they loaded, which indicates they have equally important contributions to the amount of variance accounted for by the factor.

Table 26. Univariate Regressions for Proficiency and Conduct Marks.

	(1)	(2)
VARIABLES	Proficiency_Grade	Conduct_Grade
Conduct_Grade	0.7410***	
	(0.0003)	
Proficiency_Grade		0.9522***
		(0.0004)
Constant	11.4345***	1.9765***
	(0.0147)	(0.0189)
Observations	2,039,942	2,039,942
R-squared	0.7056	0.7056
Standard errors are in parentheses.		
*** p < .01, ** p < .05, * p < .1		

D. PRACTICALITY

Practicality means that performance measures are observable, interpretable, usable, and acceptable to those who need them to make personnel decisions (Smith, 1976). Regarding observability, raters are observing performance that is predictive of future performance in terms of fitness report scores. It is quite possible that the marks could be even stronger predictors if the rating format was clearer on what specific behaviors the rater should be observing.

Proficiency and conduct marks, in terms of the rating format, are not highly interpretable. An inherent limitation of a graphic rating scale is that the rater must infer the traits of the Marine based on observed behavior (Wiese & Buckley, 1998). There is some evidence of this in results for interrater reliability. In addition, many of the attributes that are meant to guide raters in their evaluation of the Marine may be causing logical error—it is difficult to discern the difference between many of the attributes of both marks.

The marks are usable in the sense that they provide information on a Marine's performance. Factor analysis results show that among the composite score elements, proficiency and conduct marks provide the most information about the performance of an E4. However, low interrater reliability indicates that the marks may not be usable for the purpose of promotion. In addition, the difference in average marks between PMOSs could mean that the marks are not particularly useful for selection to competitive programs.

VI. CONCLUSIONS AND RECOMMENDATIONS

This study finds little direct evidence to support that proficiency and conduct marks are leading to unfair evaluations or promotions. Proficiency and conduct marks are marginally effective performance measures, and they provide at least some information on the Marine's true performance. There is evidence that marks are not consistent between raters, though the effect may be too small to affect promotions. The most significant weakness of proficiency and conduct marks is the rating format, which is likely the cause of low interrater reliability and inflation.

A. CONCLUSIONS

The study uses the hypotheses displayed in Table 27, based largely on the review of academic literature and a qualitative assessment of the marks, rating format, and information about rater training. The research questions address the measures of effectiveness: reliability, validity, and accuracy. This study includes an assessment of practicality based on the answers to all the research questions.

Table 27. Summary of Hypotheses Tested in This Study.

Research Question	Null Hypotheses—Reliability, Validity, and Accuracy	Literature Suggests…	Supporting Evidence?
R1	Pro/con marks are stable	Yes	Yes
R2	Pro/con marks are consistent between raters	No	Inconclusive
V1	Pro/con marks are important contributions to a Marine's composite score	Yes	Yes
V2	Pro/con marks predict future performance	Yes	Yes
A1	Pro/con marks differentiate between levels of performance	No	Yes
A2	Pro/con marks are not inflated	No	No
A3	Pro/con marks are distinct measures of performance	No	No

Hypotheses—Practicality	Supporting Hypotheses	Literature Suggests…	Supporting Evidence?
Raters can easily interpret the rating format	R2, A3	No	No
Raters are able to infer traits from observed behavior	V2	No	Yes
Pro/con marks are usable for the purpose of promotion decisions	R2, V1, V2, A1	Inconclusive	Inconclusive

1. Reliability

Proficiency and conduct marks appear to be stable year to year. Although this study does not find conclusive evidence to support this form of reliability, it does indicate that marks are not fluctuating randomly. More precise estimates of stability may be obtained by comparing the marks assigned to a group of Marines by the same rater at two different points in time when performance is unlikely to change. If a Marine receives different marks and the Marine's performance has not changed, then the marks have low stability.

There is evidence to support that proficiency and conduct marks are not entirely consistent between raters. Results for interrater reliability are statistically significant yet require further examination to determine practical significance. For instance, the statistically significant effect of a proficiency mark for an administrative specialist assigned to a ground unit is -0.0428. A literal interpretation is that an administrative specialist with a 4.5 proficiency mark at a non-ground unit is expected to receive a 4.4572 if assigned to a ground unit, on average. Further analysis is required to determine if the effect is large enough to influence promotion timing between equally performing Marines.

2. Validity

Of all the observed performance measures used in the composite score, proficiency and conduct marks for E4s are the most predictive of future performance in terms of fitness report scores. These results alone are encouraging. Proficiency and conduct marks are capturing, at least to some extent, the same type of performance that is recognized under a different performance evaluation system.

Exploratory factor analysis reveals the underlying performance elements in the composite score associated with each of the factors. The factors, which I call person–organization fit, physical fitness, and human capital, indicate the performance-related variables that are most important in measuring a Marine's performance. For E4s, proficiency and conduct marks, represented by the person–organization fit factor, provide the most information about a Marine's performance.

3. Accuracy

The most noteworthy concern with accuracy is that proficiency and conduct marks both measure the same performance. Barring adverse conduct, a Marine's conduct mark is predicted by his or her proficiency mark. Both marks are slightly inflated, though with no apparent harm to the promotion system. If anything, it is possible the inflation is allowing for better differentiation between performance levels.

4. Practicality

Raters are having difficulty interpreting the rating format according to the results of low interrater reliability and low accuracy in terms of the marks being distinct. The results are expected because graphic rating scales force the rater to infer a trait about someone based on observed behavior (Wiese & Buckley, 1998). Nevertheless, raters for the most part are assigning marks commensurate with the Marine's potential to perform at the next higher grade, as evidenced by predictive validity tests. Thus, there is evidence that although raters are having difficulty interpreting the rating format, they are successfully observing the performance that is intended to be evaluated.

Lastly, there are contradicting results that the marks are usable for the purpose of promotion decisions. On the one hand, the marks are predictive of future performance and align with the intent of the promotion system. On the other hand, the marks vary between rater, which could lead to unfair promotions. Further analysis of promotion timing within PMOSs across unit type is required to further support or discredit this hypothesis.

B. RECOMMENDATIONS TO IMPROVE THE EFFECTIVENESS OF PROFICIENCY AND CONDUCT MARKS

1. Keep Subjective Performance Measures and Improve the Interpretability of the Rating Format

Subjective performance measures give the performance evaluation system flexibility, as well as the ability to capture relevant performance behavior that is not practical to measure with quantitative measures. Yes, subjective performance measures are prone to biases (Landy & Farr, 1980), and biases reduce the effectiveness of the

performance evaluation system. The current system likely reduces biases, however, because the marks pass through multiple evaluators. Additionally, improvements to the rating format can increase accuracy and further reduce unintentional biases.

I recommend improving the graphic rating scale currently in use. Improving the interpretability of the format and the performance measures will increase ease of use and consistency between raters. To reduce the costs associated with adjusting related policies, I recommend that the scoring scheme remain the same and that instead of measuring proficiency or conduct, evaluators measure the attributes that define proficiency and conduct. As suggested by Landy & Farr (1983), the scale anchors for each attribute may be task- and behavior-oriented. The attribute scores combine to form an overall proficiency rating or conduct rating.

In addition to reformatting the rating scales, both marks should be redefined in order to measure behavior relevant only to proficiency or conduct. Proficiency should be redefined in order to better measure behavior relevant to performance in a Marine's specialty or primary duty. Conduct should be redefined in order to better measure a Marine's conduct as a Marine unrelated to his or her specialty or primary duty.

2. Expand on Training Given at PME Courses to Include Education Related to Cognitive Biases

Teaching raters how to avoid cognitive-related biases is one way to improve accuracy (Pursell et al., 1980). In addition to rating format improvements, training can further reduce the effect of errors such as halo, logical, central tendency, and leniency, as well as selective recall. Additionally, frame of reference training, similar to tactical decision games, could also be used. Frame of reference training allows students to work through scenarios, assign marks, and receive feedback from the instructor on how to improve their evaluation. This type of training will improve consistency between raters as well as accuracy. However, designing effective curricula requires subject matter experts to explore the specific training needs. In addition, subject matter experts first need to estimate the value of said training and determine if the costs associated with training are justified.

3. Move Proficiency and Conduct Marks into Marine Corps Order 1610.7, *Performance Evaluation System*, Instead of the *Marine Corps Individual Records Administration Manual*

Moving proficiency and conduct marks from MI, who authors and manages the IRAM, to MMRP, who authors and manages the fitness report system, will

> further allow the Marine Corps the ability to professionalize a performance evaluation continuum by combining performance evaluations under one Branch within the Manpower Management Division. However, additional costs related to expanding the role of MMRP will need to be considered such as systems integration and increases in the manpower workforce for MMRP. (R. VanOostrom, personal communication, February 17, 2017)

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. DATA CODING

Table 28. Individual Award Coding.

Variable	Award Description
High_Award_gain	AIR MEDAL-INDIVIDUAL ACTION AIR MEDAL-STRIKE/FLIGHT BRONZE STAR MEDAL DEFENSE MERITORIOUS SERVICE MEDAL DEPARTMENT OF STATE HEROISM AWARD DISTINGUISHED SERVICE MEDAL-ARMY DISTINGUISHED SERVICE MEDAL-COAST GUARD MERITORIOUS SERVICE MEDAL NASA DISTINGUISHED SERVICE MEDAL NASA EXCEPTIONAL BRAVERY MEDAL NAVY AND MARINE CORPS MEDAL NAVY CROSS PURPLE HEART SILVER STAR MEDAL
Commend_Medal_gain	AIR FORCE COMMENDATION MEDAL ARMY COMMENDATION MEDAL JOINT SERVICE COMMENDATION MEDAL NAVY AND MARINE CORPS COMMENDATION MEDAL
Achiev_Medal_gain	AIR FORCE ACHIEVEMENT MEDAL ARMY ACHIEVEMENT MEDAL JOINT SERVICE ACHIEVEMENT MEDAL NAVY AND MARINE CORPS ACHIEVEMENT MEDAL
Combat_Action_gain	COMBAT ACTION RIBBON
Low_Award_gain	CERTIFICATE OF APPRECIATION CERTIFICATE OF COMMENDATION (INDIVIDUAL AWARD) LETTER OF COMMENDATION MERITORIOUS MAST
LOA_gain	LETTER OF APPRECIATION
Volunteer_Medal_gain	MILITARY OUTSTANDING VOLUNTEER SERVICE MEDAL

Table 29. Technical and Nontechnical Categorization and Coding of PMOS.

Nontechnical				Technical							
111	481	2311	4671	321	2141	4341	6111	6176	6258	6332	6492
136	811	2621	5512	511	2146	5711	6112	6211	6276	6333	6493
231	1161	2631	5524	612	2147	5939	6113	6212	6281	6336	6499
261	1171	2651	5811	613	2148	5942	6114	6213	6282	6337	6531
311	1316	3043	5812	614	2171	5948	6116	6214	6283	6338	6541
313	1341	3051	5831	621	2671	5951	6122	6216	6286	6386	6694
331	1345	3052	6042	622	2673	5952	6123	6217	6287	6414	6821
341	1361	3112	6046	623	2674	5953	6124	6218	6288	6423	6842
351	1371	3381	6672	627	2676	5954	6132	6222	6312	6432	7011
352	1391	3432	7041	628	2821	5974	6151	6223	6313	6433	7051
411	1812	3451	7212	651	2831	5979	6152	6226	6314	6463	7234
431	1833	3521		842	2841	6048	6153	6227	6316	6466	7236
121	1834	3531		844	2844	6062	6154	6251	6317	6467	7242
151	1834	4421		847	2846	6072	6156	6252	6322	6469	7257
161	2111	4611		861	2847	6073	6172	6253	6323	6482	7314
451	2131	4612		1141	2871	6074	6173	6256	6324	6483	
471	2161	4641		1142	2887	6092	6174	6257	6326	6484	

Table 30. Unit Type Coding and Description.

Variable	Reporting Unit Code (RUC)	Examples
Unit_HQMC_Student (Unit_Nontrad)	DPI=9 plus the following: 6050 30381 30382 31301 31316 31318 31319 31340 31350 31351 31352 31353 31354 31360 31400 31401 31407 33350 33351 33352 33353 33354 33355 33808 35102 53720 54060 54061 54065 54069 54071 54078 54079 54080 54081 80222 35101	MCCDC, H&S BN HQMC HENDERSON HALL, SCHOOL OF INFANTRY (PERM PERS)
Unit_Special_Duty (Unit_Nontrad)	DPI = 16	MARINE CORPS DETACHMENT
Unit_OCONUS (Unit_Nontrad)	DPI = 27	CBTLOGREGT 37, 3D MAINT BN CBTLOGREGT 35
Unit_Logistics (Unit_Logistics)	21300 21301 21302 21303 21304 21305 21307 21308 21310 21311 21312 21313 21314 21316 21330 27012 27013 27036 27104 27110 27113 27117 27118 27119 27121 27122 27124 27125 27126 27127 27135 27139 27140 27146 27150 27151 27152 27160 27161 27162 27163 27164 27337 27340 27341 27342 27344 27350 27351 27352 27354 27360 27361 27362 27363 27366 27367 27368 27369 27371 27380 27381 27382 27383 27384 27386 27387 27388 27389 28266 28270 28271 28272 28273 28274 28280 28281 28282 28283 28284 28285 28286 28287 28288 28289 28290 28291 28292 28293 28294 28301 28303 28304 28307 28309 28310 28311 28313 28318 28319 28321 28322 28324 28325 28326 28327 28328 28333 28334 28335 28348 28349 28352 28354 28355 28357 28358 28364 28366 28367 28368 28369 28374 28375 28376 28380 28381 28382 28383 28390 28391 28392 29023 29029 29033 29034 29037 29039 29109 45614 69009	2D MAINT BN CBTLOGREGT 25 2D MLG, 8TH ENGR SPT BN 2D MLG, 2D SUPPLY BN CBTLOGREGT 25 2D MLG

Variable	Reporting Unit Code (RUC)	Examples
Unit_Infantry (Unit_Ground)	11110 11120 11130 11160 11170 11180 11210 11220 11230 12110 12111 12112 12113 12114 12115 12120 12121 12122 12124 12125 12126 12130 12140 12141 12142 12143 12144 12145 12146 12160 12161 12162 12163 12164 12165 12170 12180 12181 12182 12185 12186 12187 12210 12212 12220 12230 12240 12250 12251 12252 12253 12254 12255 13110 13120 13130 13160 13170 13210 13220 13230 13310 13311 13313 13314 13315	1ST BATTALION 3D MARINES, 2NDBN 7THMAR 1STMARDIV, 3RDBN 8THMAR 2D MARDIV
Unit_Recon (Unit_Ground)	11060 11700 11701 11702 11703 11704 11707 11708 12016 12190 12191 12192 12193 12194 12196 12197 13700 28350	1ST LIGHT ARMORED RECON BN, 2D RECONNAISSANCE BATTALION
Unit_MAG (Unit_Air)	00271 00272 00273 00274 00371 00372 00373 00374 01012 01020 01065 01068 01074 01086 01115 01121 01122 01158 01161 01162 01163 01166 01169 01171 01173 01175 01181 01185 01190 01191 01192 01194 01195 01203 01205 01211 01212 01214 01223 01224 01225 01227 01231 01232 01237 01238 01239 01251 01252 01261 01263 01264 01266 01267 01268 01269 01303 01311 01312 01314 01323 01331 01332 01352 01361 01363 01364 01365 01366 01367 01369 01461 01462 01463 01464 01465 01466 01467 01469 01513 01519 01533 01542 01561 01562 01567	HMH-361 MAG-16 3RDMAW, MALS-13 MAG-13 3D MAW, MWSS-274 MAG-29 2DMAW, VMA-231 MAG-14 2ND MAW, VMFA(AW)-225 MAG-11 3RDMAW
Unit_Tank (Unit_Ground)	21410 21411 21412 21413 21414 21415 21420 21421 21422 21423 21424 21425 21431	1ST TANK BATTALION 1STMARDIV
Unit_Arty (Unit_Ground)	11310 11311 11313 11314 11315 11320 11321 11323 11324 11325 11330 11331 11333 11334 11335 11336 11340 11341 11343 11344 11345 11346 12310 12311 12313 12314 12315 12320 12321 12323 12324 12325 12330 12331 12333 12334 12335 12350 12351 12352 12353 12354 12362 12363 12364	2D BATTALION 10TH MARINES, 5TH BATTALION 11TH MARINES 1ST MARDIV
Unit_MACG (Unit_Air)	00207 00208 00209 00219 00307 00308 00309 00311 00820 00830 00840 00842 00843 00852 00853 00870 00871 00872 00873 00877 00880 00881 00882 00883 00887 00920 00921 00922 00923 00924	MACS-1 MACG-38, MWCS-28 MACG-28 2D MAW, VMU-3 MACG-38 3D MAW, MASS-1

Variable	Reporting Unit Code (RUC)	Examples
	00930 01144 01145 01480 01490 01495	MACG-28 2D MAW
Unit_Engineer (Unit_Ground)	11400 11401 11403 11404 11405 11406 11407 12400 12401 12403 12404 12405 12407 12408 13420 13421 13422 13423 13424	1ST CBTENGR BN 1STMARDIV, 2D CBT ENGR BN 2D MARDIV
HQ_Logistics	28370 28371 31001 38440 38441 38445 45020 27100 27101 27102 27103 27105 27108 27370 28300 28302 28305 28306 29016	CBTLOGREGT 27 2D MLG, CBTLOGREGT 1
HQ_Ground	02300 11000 11001 11100 11104 11154 11190 11200 11204 11300 11303 12000 12001 12100 12101 12150 12151 12201 12290 12300 12301 13100 13101 20021 20034 20080 20132 20146 20149 20151 20171 20173 20176 20177 20179 20180 20181 20199 20251 20310 20361 20362 20371 20372 20373 20420 35010 45683 20197 20198	1ST INTELLIGENCE BATTALION, II MEF COMMAND ELEMENT, HEADQUARTERS BATTALION
HQ_Air	00011 00013 00014 00016 00026 00029 00031 00039 00044 00045 00072 00073 00201 00202 00300 00376 01053 01070 01071 01075 01079 01081 01243 01510 02001 02021 02030 02031 02200 02201 02208 02230 02231 02303 45644	HQ MAG-31 2NDMAW, MACG-38 3D MAW, MWHS-2 2D MAW
Consolidated unit type categories are in parentheses.		

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. ADDITIONAL FACTOR ANALYSIS RESULTS

Table 31. Factor Loading Comparisons between 3-Factor and 2-Factor Models.

	3-factor model			2-factor model	
Variable	Factor1	Factor2	Factor3	Factor1	Factor2
E3 Technical PMOS					
Composite_Rifle					
Composite_PFT			0.5764		0.4111
Composite_CFT			0.5951		0.3584
Composite_Pro		0.8204			0.8221
Composite_Con		0.8432			0.7944
Composite_TIG	0.9252			0.9262	
Composite_TIS	0.9202			0.927	
Composite_Educ_Bonus	0.3744			0.3707	
Composite_Duty_Bonus					
Composite_Recruiting_Bonus					
E3 Nontechnical PMOS					
Composite_Rifle					
Composite_PFT			0.5608		0.3035
Composite_CFT			0.5799		
Composite_Pro		0.8348			0.8445
Composite_Con		0.8447			0.818
Composite_TIG	0.927			0.9226	
Composite_TIS	0.9199			0.9236	
Composite_Educ_Bonus					
Composite_Duty_Bonus					
Composite_Recruiting_Bonus					
E4 Technical PMOS					
Composite_Rifle					
Composite_PFT		0.6104		0.3523	
Composite_CFT		0.5973			-0.3172
Composite_Pro	0.8779			0.8673	
Composite_Con	0.8923			0.853	
Composite_TIG			0.7206		0.6811
Composite_TIS			0.6102		0.6797
Composite_Educ_Bonus			0.3237	0.3398	
Composite_Duty_Bonus					
Composite_Recruiting_Bonus					

Variable	3-factor model			2-factor model	
	Factor1	Factor2	Factor3	Factor1	Factor2
E4 Nontechnical PMOS					
Composite_Rifle					
Composite_PFT			0.5861	0.3283	
Composite_CFT			0.5948		
Composite_Pro	0.8822			0.8814	
Composite_Con	0.8983			0.8653	
Composite_TIG		0.7936			0.7723
Composite_TIS		0.7076			0.7271
Composite_Educ_Bonus		0.3218		0.343	
Composite_Duty_Bonus					
Composite_Recruiting_Bonus					

Blanks represent values less than .3.

LIST OF REFERENCES

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(3), 20–25. Retrieved from https://jalt.org/test/PDF/Brown31.pdf

Cederblom, D., & Pemerl, D. E. (2002). From performance appraisal to performance management: One agency's experience. *Public Personnel Management, 31*(2), 131–140. Retrieved from http://libproxy.nps.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6873752&site=ehost-live&scope=site

Center for Naval Analyses. (2015). Appendix B: Active component enlisted accessions, enlisted force, officer accessions, and officer corps tables. Retrieved from https://www.cna.org/pop-rep/2014/summary/summary.html

Chung, G. K., Nagashima, S. O., Delacruz, G. C., Lee, J. J., Wainess, R., & Baker, E. L. (2011). *Review of rifle marksmanship training research* (CRESST Report 783). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://hdl.handle.net/10945/37927

Clemens, A., Malone, L., Phillips, S., & Lee, G. (2012). *An evaluation of the fitness report system for marine officers*. Alexandria, VA: Center for Naval Analyses.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*(1), 130–135. Retrieved from http://web.a.ebscohost.com.libproxy.nps.edu/ehost/pdfviewer/pdfviewer?sid=a14c65f0-7cbf-4edc-afcc-95e3fff22e2c%40sessionmgr4010&vid=1&hid=4104

Cole, A. (2014). *U.S. Marine Corps enlisted retention: An analysis of stakeholder incentives for the retention of tier 1 first-term Marines* (Master's thesis). Monterey, CA: Naval Postgraduate School. Retrieved from http://hdl.handle.net/10945/41360

Crider, L. (2015). *Effectiveness of the United States Marine Corps tiered evaluation system* (Master's thesis). Monterey, CA: Naval Postgraduate School. Retrieved from http://hdl.handle.net/10945/45175

Doeringer, P. B., & Piore, M. J. (1985). *Internal labor markets and manpower analysis.* Armonk, NY: M.E. Sharpe.

Dunford, J. F. (2015). *U.S. Marine Corps 36th Commandant's planning guidance.* Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900391/36th-commandants-planning-guidance/

Griner, M. S. (2016). *Quality of USMC officers: Buildup vs. reduction in forces* (Master's thesis). Monterey, CA: Naval Postgraduate School. Retrieved from http://hdl.handle.net/10945/48529

Hosek, J. R., & Mattock, M. G. (2003). *Learning about quality: How the quality of military personnel is revealed over time* (No. RAND/MR-1593-OSD). Santa Monica, CA: RAND. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA411795

Hutchinson, S. (2013). *Performance management.* London, UK: CIPD. Retrieved from http://site.ebrary.com/lib/nps/Top?id=11161957

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33*(3), 595–640. Retrieved from http://web.a.ebscohost.com.libproxy.nps.edu/ehost/pdfviewer/pdfviewer?sid=124cc833-6231-4cdb-b78f-c00929355ab1%40sessionmgr4009&vid=2&hid=4104

Keller, G. (2009). *Statistics for management and economics* (8th ed.). Mason, OH: South-Western Cengage Learning.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72–107. doi://dx.doi.org/10.1037/0033-2909.87.1.72

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York, NY: Academic Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. Retrieved from http://web.a.ebscohost.com.libproxy.nps.edu/ehost/pdfviewer/pdfviewer?sid=964601c3-213c-47f9-817d-97bf8a9ad614%40sessionmgr4006&vid=1&hid=4104

Lazear, E. P., & Gibbs, M. (2015). *Personnel economics in practice* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Longenecker, C. O., Liverpool, P. R., & Wilson, K. Y. (1988). An assessment of manager/subordinate perceptions of performance appraisal effectiveness. *Journal of Business & Psychology*, *2*(4), 311–320. Retrieved from http://libproxy.nps.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16853433&site=ehost-live&scope=site

MacDonald, H. A., & Sulsky, L. M. (2009). Rating formats and rater training redux: A context-specific approach for enhancing the effectiveness of performance management. *Canadian Journal of Behavioural Science, 41*(4), 227–240. Retrieved from http://search.proquest.com.libproxy.nps.edu/docview/220509061/abstract/F9D091F27E3D4DFFPQ/1

Marine Corps University. (2012). *Proficiency and conduct marks: Student guide. United States Marine Corps Enlisted Professional Military Education Corporals Course* (CPL-ADMIN-2442A). Retrieved from https://vcepub.tecom.usmc.mil

Mayberry, P. W. (1986). *Confirming differences in relative-value proficiency marks*. Arlington, VA: Center for Naval Analyses. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD= ADA175362

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criterion. *Personnel Psychology, 33*(1), 91–96.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428. doi://dx.doi.org/10.1037/0033-2909.88.2.413

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago, IL: Rand McNally.

StataCorp. (2013). Factor postestimation. In *Stata 13 multivariate statistics reference manual*. College Station, TX: Stata Press. Retrieved from http://www.stata.com/manuals14/mvfactorpostestimation.pdf

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*(4), 501–510.

United States Marine Corps. (2000). *Marine Corps individual records administration manual (short title: IRAM)* (Marine Corps Order P1070.12K). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/899281/mco-p107012k-wch-1/

United States Marine Corps. (2001). *Selecting, screening, and preparing enlisted Marines for special duty assignments and independent duties* (Marine Corps Order P1326.6D Ch 1). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900162/mco-p13266d-w-ch-2/

United States Marine Corps. (2007). *Marine Corps combat marksmanship programs (MCCMP)* (Marine Corps Order 3574.2K). Retrieved from http://www.marines.mil/Portals/59/MCO%203574.2K.pdf

United States Marine Corps. (2008). *Marine Corps physical fitness program* (Marine Corps Order 6100.13 W/CH 1). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900398/mco-610013-wch-2/

United States Marine Corps. (2010). *Enlisted retention and career development program* (Marine Corps Order 1040.31). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/899418/mco-104031/

United States Marine Corps. (2012). *Marine Corps promotion manual, volume 2, enlisted promotions* (Marine Corps Order P1400.32D Ch 2). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/899517/mco-p140032d-wch-1-2/

United States Marine Corps. (2013). *Separation and retirement manual* (Marine Corps Order 1900.16). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900480/mco-190016-wch-1/

United States Marine Corps. (2014). *Marine Corps combat marksmanship programs (MCCMP)* (Marine Corps Order 3574.2L). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900336/mco-35742l/

United States Marine Corps. (2015a). *Enlisted to officer commissioning programs* (Marine Corps Order 1040.43B). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/898711/mco-104043b/

United States Marine Corps. (2015b). *Performance evaluation system (short title: PES)* (Marine Corps Order 1610.7). Retrieved from http://www.marines.mil/News/Publications/ELECTRONIC-LIBRARY/Electronic-Library-Display/Article/900397/mco-16107/

United States Marine Corps. (2016a). *Changes to the physical fitness test (PFT), combat fitness test (CFT), and body composition program (BCP)* (All Marine Message 022/16). Retrieved from http://www.fitness.marines.mil/almar/

United States Marine Corps. (2016b). *FY17 first term alignment plan (FTAP) quarterly assessment* (Marine Administration Message 540/16). Retrieved from http://www.marines.mil/News/Messages/Messages-Display/Article/971326/fy17-first-term-alignment-plan-ftap-quarterly-assessment/

Wiese, D. S., & Buckley, R. M. (1998). The evolution of the performance appraisal process. *Journal of Management History (Archive), 4*(3), 233–249. doi:10.1108/13552529810231003

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California